Acknowledgments

First of all, I would like to thank my supervisor doctor Natércia Brás for having accompanied me throughout the course of this work, as well as all the work that came before it. Even though she had a child, she always tried to be available and answer all the questions I had, as well as suggest new ways to solve problems that did not occur to me.

I would also like to show my gratitude to my co-supervisor, Professor Maria João Ramos, for giving me the opportunity to work in such a great group and for the ideas given when all the possibilities seemed to be exhausted. I would also like to thank Professor Pedro Alexandrino Fernandes for also giving me the opportunity to work in this group and for the input and suggestions given, which also helped a great deal in moving the work forward.

To all the people in the group, I would like to thank not only for the work-related suggestions and ideas, which also opened my eyes to new possibilities and hypotheses, but also for the excellent work environment, which makes it a pleasure working with them. I am very thankful to everyone for making every day interesting from start to finish.

I would like to thank all the people who supported me during the course of this work. In particular, I thank Gaspar for introducing me to this area, for helping me whenever I needed and for distracting me from my work and letting me distract him from his. I also want to thank Diana for all the support and friendship shown.

I am very thankful for my parents, for all the unconditional support they have always given me and without whom this project would not have been possible. I also appreciate the support given by my sister, who has accompanied me on every step since I was little.

Lastly, I am very grateful for Inês, who has been by my side the entire time, always showing me support and keeping me motivated when I most needed it. The encouragement to work, as well as all the advices and distractions were essential for me to conclude this work, especially during the later stages.

Abstract

Herpes Simplex Virus type I (HSV-1) is a virus which generally infects non-dividing cells. In order for HSV to replicate, the glycoproteins in its envelope must interact with host cellular receptors, leading to the fusion of the envelope with the cell membrane and consequent delivery of the capsid and DNA into the cytoplasm.

Heparan sulfates (HS) are linear polysaccharides that consist of repeating disaccharide units of uronic acid-(1-4)-D-glucosamine, with different patterns of sulfation and acetylation. HS are involved in several biological interactions, such as assisting viral infection or regulating blood coagulation. As the HS chain is being formed, it suffers several modification processes which affect its biological function.

3-*O*-sulfotransferase (3-OST) is one of the enzymes involved in the HS biosynthesis. It is responsible for the transfer of a sulfo group to glucosamine units linked to uronic acid residues. Different 3-OST isoforms have distinct substrate specificities and produce HS with different biological functions, with isoform 3 conferring HSV-1 entry receptor activity to HS. 2-*O*-sulfotransferase (2-OST) is responsible for the step previous to the one catalyzed by 3-OST, in the HS biosynthesis. It catalyzes the transfer of a sulfate group to the 2-OH position of IdoA or GlcA, adjacent to glucosamines, in the HS chain.

The main goal of this work is to determine the catalytic mechanism of 3-OST and 2-OST with atomic detail and to find potential 3-OST inhibitors, using computational methods. In order to achieve this, we applied the ONIOM method, to a large enzyme model divided into two regions. The protonation state of key residues was assessed through Molecular Dynamics (MD) simulations, as well as pK_a estimations using web-based prediction tools. The MD simulations also served to relax the systems being studied. A virtual screening study was performed, in an in-house developed library of compounds, in order to find compounds with potential 3-OST inhibitor activity.

Our results show that the molecular mechanisms of 3-OST and 2-OST occur by a single mechanistic step. We were able to determine the protonation state of the key amino acids in the active site. We were also able to obtain optimized reactants, transition state (TS) and products geometries for 3-OST and unoptimized ones for 2-OST, as well as activation and reaction free energies for both enzymes. We identified two compounds, phomoidride B and Barceloneic acid A, as potential 3-OST inhibitors.

Keywords Sulfotransferases, Herpes Simplex Virus, Heparan Sulfates, QM/MM, MD simulations, Virtual Screening, 3-O-sulfotransferase, 2-O-sulfotransferase

Resumo

O vírus do Herpes Simplex tipo I (HSV-1) é um vírus que infeta, geralmente, células que não estão em divisão. Para este se replicar, as glicoproteínas na sua cápsula interatuam com recetores celulares do hospedeiro, levando à fusão da cápsula com a membrana celular e consequente transferência da cápside e DNA para o citoplasma.

Heparano Sulfatos (HS) são polissacarídeos lineares, que consistem em unidades dissacarídeas repetidas de ácido urónico-(1-4)-D-glucosamina, com diferentes padrões de sulfatação e acetilação. Os HS estão envolvidos em diversas interações biológicas, tais como a regulação da coagulação sanguínea e assistência à entrada de vírus.

A 3-OST é uma das enzimas envolvida na biossíntese de HS. É responsável pela transferência de um grupo sulfo para unidades de glucosamina, ligadas a unidades de ácido urónico. Isoformas diferentes têm diferentes especificidades para substrato, produzindo HS com diferentes funções biológicas, sendo que a isoforma 3 confere ao HS actividade de recetor de entrada do HSV-1. A 2-OST é responsável pelo passo anterior ao catalisado pela 3-OST, na síntese de HS. Catalisa a transferência de um grupo sulfato para a posição 2-OH de unidades de IdoA ou GlcA na cadeia de HS.

O objetivo deste trabalho é determinar o mecanismo catalítico da 3-OST e 2-OST com detalhe atomístico e determinar inibidores da 3-OST, utilizando métodos computacionais. Para tal, aplicamos o método ONIOM a um modelo enzimático dividido em duas regiões. O estado de protonação dos resíduos chave foi avaliado através de simulações de Dinâmica Molecular (MD), bem como através da utilização de ferramentas de previsão de pK_as online. As simulações também tiveram como objetivo o relaxamento dos sistemas. Um estudo de screening virtual foi efetuado, de forma a encontrar compostos com potencial atividade inibitória para a 3-OST.

Os resultados mostram que os mecanismos moleculares da 3-OST e 2-OST ocorrem através de um único passo mecanístico. Foi possível determinar o estado de protonação dos aminoácidos chave no centro ativo. Foi possível obter geometrias otimizadas de reagentes, estados de transição (TS) e produtos para a 3-OST e geometrias não otimizadas para a 2-OST, bem como energias de ativação e reação para ambas. Foram identificados dois compostos, o phomoidride B e o ácido Barceloneico A como potenciais inibidores da 3-OST.

Palavras-chave Sulfotransferases, Vírus do Herpes Simplex, Heparano Sulfatos, QM/MM, Simulações MD, Screening Virtual, 3-O-sulfotransferase, 2-O-sulfotransferase

Index

Figure Index

Table Index

List of Abbreviations

2-OST - 2-*O*-sulfotransferase

3-OST - 3-*O*-sulfotransferase

6-OST - 6-*O*-sulfotransferase

ADMET - Absorption, Distribution, Metabolism, Excretion and Toxicity

AIDS - Acquired Immunodeficiency Syndrome

B3 - Three parameter Becke functional

B3LYP - Combination of three parameter Becke functional and LYP correlation functional

BLYP - Combination of Becke gradient exchange functional with LYP gradient correlation functional

CGTO - Contracted Gaussian-type Orbitals

CMV - cytomegalovirus

DFT - Density Functional Theory

EBV - Epstein-Barr virus

EE - Electronic Embedding

gaff - General AMBER force field

GAG - Glucosaminoglycan

GalT-1 - UDP-galactosyltransferase I

GalT-2 - UDP-galactosyltransferase II

GGA - Generalized Gradient Approximation

GlcA - UDP-glucuronic acid

GlcAT-1 - UDP-glucuronic acid transferase I

GlcNAc - N-acetylglucosamine

GlcNS - N-sulfoglucosamine

GTO - Gaussian-type Orbitals

HEG - Homogenous Electron Gas

HHV-6 - Herpesvirus type 6

HHV-7 - Herpesvirus type 7

HS - Heparan Sulfates

HSPG - Heparan Sulfate Proteoglycans

HSV-1 - Herpes Simplex Virus type I

HSV-2 - Herpes Simplex Virus type II

IdoA - Iduronic acid

IdoA2S - 2-*O*-sulfated iduronic acid

IMOMM - Integrated MO + MM method

IMOMO - Integrated MO + MO method

Kaposi's sacroma-associated herpesvirus (KSHV)

L(S)DA - Local (Spin) Density Approximation

LYP - Lee, Yang and Parr

MD - Molecular Dynamics

ME - Mechanical Embedding

MM - Molecular Mechanics

MO - Molecular Orbitals

NDST - N-deacetylase/N-sulfotransferase

NDST-1 - N-deacetylase/N-sulfotransferase-1

ONIOM - "Our Own n-layered Integrated MO and MM" method

PAP - 3'-phosphoadenosine 5'-phosphate

PAPS - 3'-phosphoadenosine 5'-phosphosulfate

parm03 - AMBER 2003 force field

PB - 3'-phosphate-binding

PDB - Protein Data Bank

PES - Potential Energy Surface

PME - Particle-Mesh Ewald

PSB - 5'-phosphosulfate-binding

QM - Quantum Mechanics

RESP - Restrained Electrostatic Potential

RMSD - Root Mean Square Deviations

SCF - Self Consistent Field

SLH3 - β-strand-loop-α-helix

ST - sulfotransferase

STO - Slater-type Orbitals

SZ - Single-ζ functions

TS - Transition State

VWN3 - Local-density approximation to the correlation functional for B3LYP by Vosco, Wilk and Nusair

VZV - varicella-zoster virus

XYLT-1 - UDP-xylosyltransferase I

XYLT-2 - UDP-xylosyltransferase II

ZDO - Zero Differential Overlap

ZPE - Zero Point Energy

β-AST-IV - β-arylsulfotransferase-IV

I. Introduction

1. Herpes Simplex Virus type I

Herpes Simplex Virus type I (HSV-1) is a part of the herpesvirus group, included in the Herpesviridae family. The herpesvirus group consists of large, enveloped, double-stranded DNA viruses, which infect both animals and humans. Included in this group are two herpes simplex viruses (HSV-1 and HSV-2), as well as Epstein-Barr virus (EBV), cytomegalovirus (CMV), varicella-zoster virus (VZV), Kaposi's sarcoma-associated herpesvirus (KSHV) and herpesvirus types 6 and 7 (HHV-6 and HHV-7), all of which infect humans.

Herpesviruses generally infect non-dividing cells and, therefore, need to synthesize their own DNA synthesis enzymes. This need leads them to have a large genome, which codes for approximately 75 viral proteins. Differences in genomic sequences between members of the herpesvirus group, which result in structural differences, can be assessed through antigenic analysis.

HSV has the widest range of cell tropism, replicating in animal and human host cells. However, its symptoms only manifest in humans. Infection by this family of viruses consists of two different stages. In the first stage after infection, the genome of the virus is present in the host cells but the virus is not activated. After this latent infection stage, reactivation of the virus may occur due to several factors, such as a compromise of the host immune system, potentially leading to severe disease. However, the mechanisms through which a latent infection evolves into an acute one are presently unknown.

In order for HSV to replicate, similarly to the other members in its family, the glycoproteins in its envelope must interact with host cellular receptors, leading to the fusion of the envelope with the cell membrane and consequent delivery of the capsid and DNA into the cytoplasm. After migration into the nucleus, the DNA is circularized. The transcription of the genome is coordinated and in a specific order, with different classes of mRNAs being produced at different points in time. The virus envelope is obtained from the host cell inner lamella of the nuclear membrane. After the virion particle is complete, it enters the cytoplasm in order to be released through the endoplasmic reticulum. The host cell metabolism is shut down, leading to death of the cell.

Although they share similarities, HSV-1 has several differences, when compared to its closest sibling, HSV-2. Both are, as previously mentioned, double-stranded DNA

viruses; however, their genome only has approximately 50% base sequence homology. Both share surface antigens, such as glycoproteins and other structural polypeptides; however, they present differences in glycoprotein gB which allows a distinction between both to be made. HSV-1 and HSV-2 diseases manifest through recurrent lesions in the skin and mucous membranes; however, type 1 prevails in the ectoderm (skin, mouth, conjuctiva, nervous system), whereas type 2 is associated with genital infection. Severe manifestations of the disease may occur, namely through infection of the eye, leading to damage to the cornea and blindness. Rare cases of HSV-1 infection may lead to encephalitis, affecting one temporal lobe and leading to cerebral edema and focal neurological signs. The virus may also be transmitted from mother to child, most commonly during delivery, leading to the development of neonatal herpes, which may result in neurological damage to the child or, ultimately, death, as observed in the majority of cases.

Herpes simplex viruses have a worldwide distribution, with humans potentially being their only host systems. They spread through direct contact with infected secretions and have varying degrees of prevalence, depending on factors like age or socioeconomic status. It is estimated that a significantly large percentage of the population has antibodies for HSV-1 [1].

Up until the point of writing, several antiviral drugs for inhibition of HSV have been developed, among which the widely distributed aciclovir [2], which is a nucleoside analog that selectively inhibits DNA replication of herpes viruses, while showing a low host cell toxicity. In infected cells, this compound is phosphorylated to a monophosphate by a viral thymidine kinase; the resulting compound is converted into the active derivative aciclovir triphosphate by host cell enzymes, which shows a potent inhibitory activity of the viral DNA polymerase, thus preventing viral replication. Aciclovir has been shown to significantly decrease the duration of the primary infection; however, it is unable to eradicate the virus in its latent form. After the release of aciclovir into the market, reports of aciclovir-resistant strains of HSV, which have mutations in the thymidine kinase gene, leading to little to no production of the enzyme, started to surface [3, 4], particularly among immunocompromised individuals, namely those with acquired immunodeficiency syndrome (AIDS) [5]. This lead to the development of new antiviral drugs which target aciclovir-resistant HSV, such as Foscarnet [6]. More recently, aciclovir derivatives, such as valacyclovir [7] and famciclovir [8, 9] have been used for the treatment of recurrent HSV genital infection. The former is a prodrug with better bioavailability,

when compared to its parent compound, which is rapidly converted to aciclovir. The latter is the prodrug of penciclovir [10], which is another nucleoside analog.

2. Heparan Sulfates

Heparan Sulfates (HS) are linear polysaccharides closely related to heparin, which consist of repeating disaccharide units of uronic acid-(1-4)-D-glucosamine. In addition, the different disaccharide units can have variable patterns of N-sulfation, O-sulfation and N-acetylation, which result in a large number of complex sequences. HS and heparin are members of the glucosaminoglycan (GAG) family of polysaccharides, along with chondroitin sulfate, dermatan sulfate and keratan sulfate [11].

HS chains are normally attached to a core protein through a serine residue to form heparan sulfate proteoglycans (HSPG). These are located at the cell surface and in the extracellular matrix, and are expressed in and secreted by most of the mammalian cells, unlike heparin, which occurs only in connective-tissue type mast cells. Proteoglycans are glycoproteins with one or more GAG chains attached to the core protein. These include the syndecan family (transmembrane proteins), the glypican family (proteins attached to the cell membrane through a glycosylphosphatidylinositol chain) and extracellular matrix proteins such as perlecan [12]. The type and amount of HS may be influenced by the amount of core protein or its competition with HS for enzymes involved in HS biosynthesis, in the Golgi compartment [13].

HS was originally discovered as an impurity in heparin in 1958, and thus named heparitin sulfate [14]. However, since its discovery, it has been found to be of critical importance to several different biological processes.

HS biosynthesis occurs mainly in the Golgi apparatus, mediated by several biosynthetic enzymes, which, with the exception of 3-*O*-sulfotransferase (3-OST), are transmembrane proteins exerting their catalytic activities inside the Golgi compartment, and begins with the assembly of the GAG chains. This assembly is initiated by the formation of a GAG-protein linkage region consisting of β-xylose, β(1-4)-galactose, β(1-3)-galactose and β(1-3)-glucuronic acid, common to heparin, chondroitin sulfate and dermatan sulfate. The linkage tetrasaccharide is formed by stepwise addition of each sugar residue. In the first reaction, the UDP-xylosyltransferases (XYLT-1 and XYLT-2) add a xylose residue in the β-anomeric configuration to a specific serine residue of the

core protein, next to a glycine flanked by acidic and hydrophobic amino acids [15, 16]. This bond is formed between an oxygen atom from the serine and the anomeric carbon (C1) of the xylose residue. Then, UDP-galactosyltransferase I (GalT-1) catalyzes the addition of a galactose residue, β(1-4)-linked to the xylose residue. Afterwards, another galactose is added, this time catalyzed by UDP-galactosyltransferase II (GalT-2), due to the increase in the chain complexity and its effect on the enzyme-substrate specificity, which results in a β(1-3) bond instead of the β(1-4) bond observed in the previously added galactose. Finally, UDP-glucuronic acid transferase I (GlcAT-1) catalyzes the addition of a UDP-glucuronic acid (GlcA) residue through a β(1-3) bond. This linkage region can be modified, namely by phosphorylation of the xylose unit and sulfation of either galactose unit, which may influence the activity of polymerization enzymes downstream [17]. The formation of the linkage region can be seen on Figure 1.

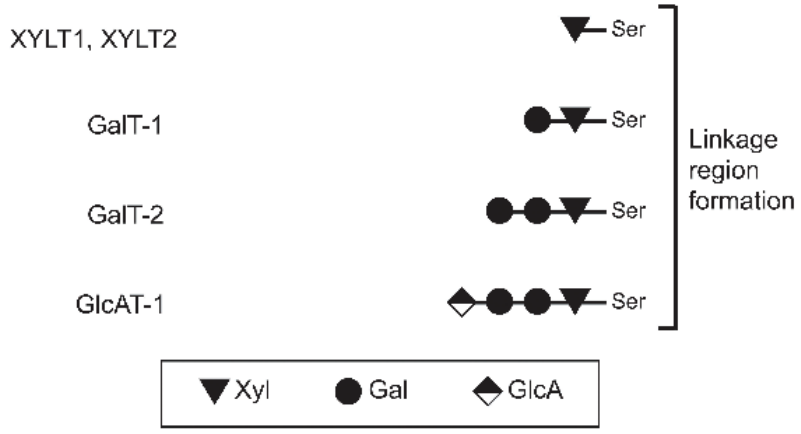

Figure 1 - Representation of the HS chain initiation. A xylose residue is added to specific serine residues in HSPG core proteins, followed by the formation of a linkage region, glucuronic acid-galactose-galactose-xylose. Adapted from [18].

HS chain elongation is mediated by two polymerization enzymes: GlcNAc transferase II that adds a α(1-4)-linked N-acetylglucosamine (GlcNAc) residue to the HS chain, and GlcA transferase II, which adds a β(1-4)-linked GlcA residue. These enzymes are encoded by the tumor suppressing genes EXT1 and EXT2. Expression of both genes is necessary for the polymerization of the HS chain to occur, and elongation is decreased when either of the genes is not expressed. The levels of EXTL3, a gene of the EXT family that is similar to the EXT3 gene, also affect HS chain elongation and size. It has been shown that this gene is responsible for the chain initiation [19, 20]. The sulfation of the HS backbone, which is mentioned below, stimulates polymerization and may also be related to the length of the HS chain [21]. This step of the HS biosynthesis can be seen on Figure 2.

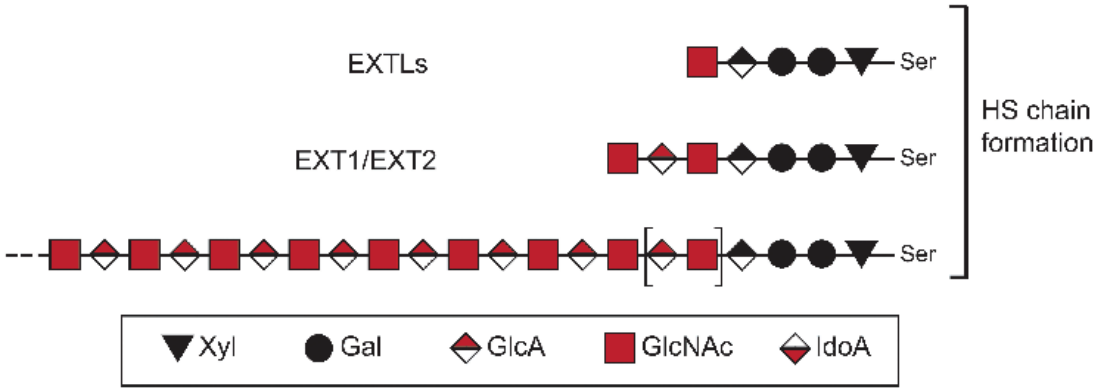

Figure 2 - Representation of the HS chain formation and elongation. EXTL3 attaches the first N-acetylglucosamine residue and EXT1/EXT2 alternately add GlcA and GlcNAc residues to the nascent chain. Adapted from [18].

As the HS chain is being formed, it can be modified differently and in diverse positions. This process is not random, it occurs in a specific and orderly manner and the modifications at each step affect the substrate of the enzyme that performs the next modification [22]. First, specific GlcNAc residues suffer N-deacetylation, which generates free amino-groups, and are then N-sulfated. Both reactions are catalyzed by the bifunctional enzyme N-deacetylase/N-sulfotransferase (NDST). Subsequent HS chain modifications depend on the presence of N-sulfoglucosamine (GlcNS) residues obtained from the modifications catalyzed by NDST. This enzyme is, therefore, responsible for the overall design of the polysaccharide [23]. After N-sulfation, most of the GlcA units positioned next to glucosamine units are converted to iduronic acid (IdoA) by C5-epimerase. This enzyme cannot convert GlcA units that are surrounded by IdoA residues already sulfated at the C2 position or by glucosamine residues that have been O-sulfated in the C6 position, which suggests that this enzyme acts before other modifications in the HS chains. After C5-epimerization, 2-*O*-sulfotransferase (2-OST) catalyzes the transfer of sulfate to the C2-position of IdoA (2-*O*-sulfation) and, with a lower activity, to GlcA in the HS chains. 2-OST presence is necessary for the stability and translocation of C5-epimerase to the Golgi apparatus. These two enzymes co-localize in the Golgi apparatus and may form functional complexes [24].

In addition to 2-*O*-sulfation, HS also suffers 6-*O*-sulfation and 3-*O*-sulfation. Unlike 2-OST, which only has one isoform, 6-*O*-sulfotransferase (6-OST) has 3 isoforms and 3-*O*-sulfotransferase (3-OST) has seven. The evolution and preservation of these 10 *O*-sulfotransferases suggests that the functionally relevant regulation of the HS biosynthesis occurs at the 6-*O*- and 3-*O*-sulfation. The modification steps detailed here can be seen on Figure 3.

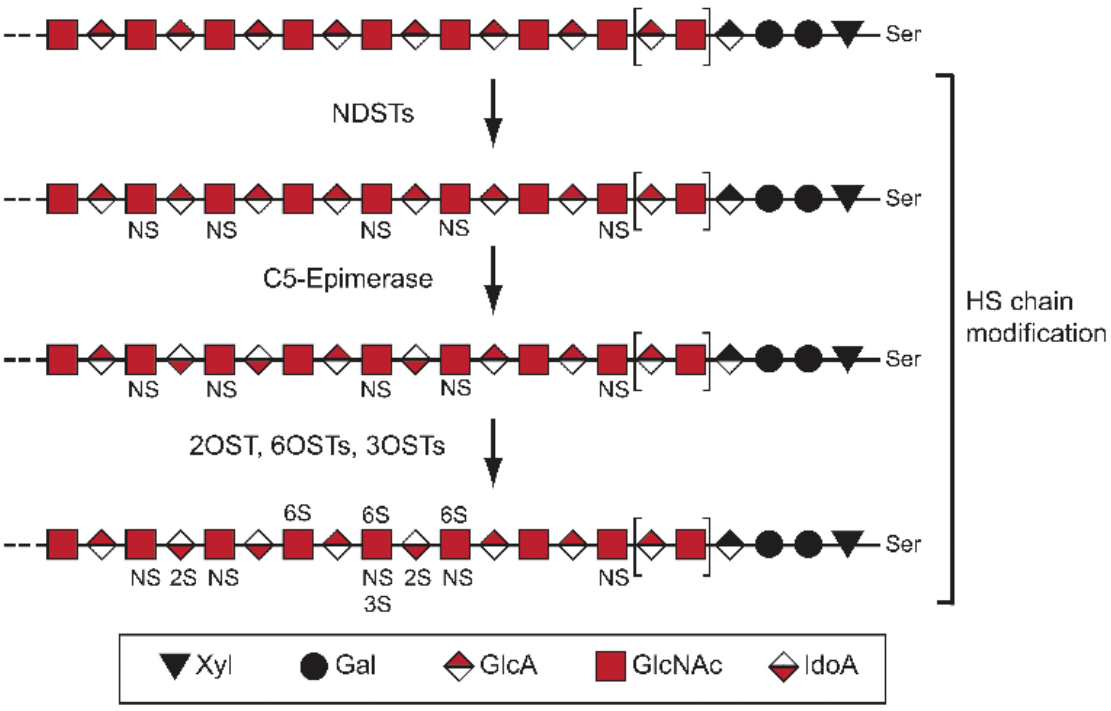

Figure 3 - Representation of HS chain modification. The chains undergo a series of processing reactions: first, NDSTs catalyze the removal of acetyl groups from GlcNAc residues and the substitution of the free amino groups with sulfate. Then, GlcA residues adjacent to GlcNS units are C5-epimerized to IdoA. The chains then suffer *O*-sulfation: IdoA units (and, less frequently, GlcA units) are sulfated at C2, reaction catalyzed by 2-OST. GlcNS units (and, less frequently, GlcNAc units) are sulfated at C6, reaction catalyzed by 6-OST. 3-OST catalyzes sulfate addition at C3 of glucosamine units, either N-sulfated or N-unsubstituted. Adapted from [18].

The substrate specificities of different 6-OST isoforms differ. The three isoforms may sulfate both GlcNAc and GlcNS residues, but sulfation occurs preferentially on GlcNS flanked by 2-O-sulfated IdoA units [25].

HS modification in the C3 position of glucosamine units is catalyzed by the 3-OST family. This modification is the rarest and is responsible for 0.5% of sulfate in the HS chains. Each 3-OST isoform transfers a sulfate group to a glucosamine unit linked at the non-reducing end to either a GlcA/IdoA unit, in the isoform 1 case, a 2-*O*-sulfated iduronic acid (IdoA2S) unit, in the isoform 3 case, or a combination of both, in the case of isoform 5. HS modified by different 3-OST isoforms has different biological activity: when it is modified by isoform 1, bound to antithrombin, it regulates the blood coagulation cascade, having intrinsic anticoagulant activity; when it is modified by isoform 3 it is bound by herpes simplex virus type I (HSV-I) glycoprotein D, serving as an entry receptor for the virus; when it is modified by isoform 5 it has both anticoagulant activity and promotes HSV-1 entry. HS modified by isoforms 2, 4 and 6 have been shown to have a biological activity similar to isoform 3 [26-28]. Biological activity of HS modified by isoform 7 has not yet been determined.

Degradation of HS occurs at the extracellular/surface level or at the intracellular level, in which case it may be recycled. In the first case, endosulfatases SULF1 and SULF2 can remove 6-*O*-sulfate groups from HS [29] and secreted heparanase can trigger the release of small, bioactive HS fragments, from HS chains attached to core proteins, leading to its degradation [30]. Intracellularly, recycling occurs by internalization of the core protein and, after partial degradation of the HS chains (coupled with the biosynthesis of new HS), migration of the core protein with the newly formed HS chains back to the cell surface [31]. Exoenzymes in the lysosomes are responsible for the irreversible intracellular HS degradation [32].

3. Sulfotransferases

Generally, sulfotransferases (STs) (E.C. 2.8.2) are enzymes that catalyze the transfer of a sulfate group from a donor molecule, normally 3'-phosphoadenosine 5'-phosphosulfate (PAPS) [33] to several different types of amine and hydroxy substrates. STs can be divided in two different classes: cytosolic STs and membrane-associated STs. The former add sulfate groups to both endogenous and exogenous small compounds, such as hormones, bioamines, drugs and xenobiotic agents. The latter are responsible for the sulfation of larger biomolecules, such as carbohydrates and proteins. The general reaction catalyzed by sulfotransferases can be seen on Figure 4.

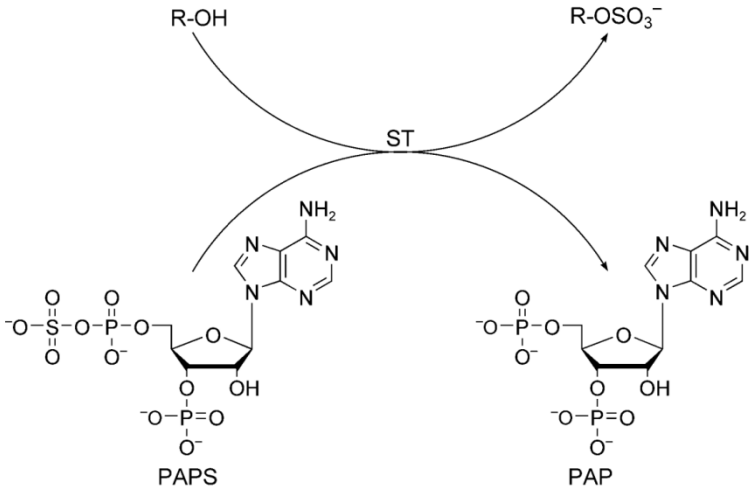

Figure 4 - General sulfotransferase-catalyzed reaction with PAPS as the co-substrate.

Cytosolic STs, contrary to what was initially assumed, are not only involved in detoxification, having important roles in the regulation of hormones and carcinogens, by forming sulfate conjugates.

Membrane-associated STs are involved in a multitude of important biological processes, ranging from viral entry and anticoagulation to leukocyte adhesion. They are responsible for the triggering of molecular recognition and signal transduction, through sulfation of proteoglycans.

This class of enzymes presents a highly conserved spherical structure, containing a single α/β fold consisting of a central parallel β sheet (four or five-stranded), surrounded by α helices. Both the structure and the amino acids are conserved for the PAPS-binding region. This region consists, generally, of a 5'-phosphosulfate-binding (PSB) loop, a 3'-phosphate-binding (PB) loop and a β-strand-loop-α-helix (SLH3) motif [34]. The largest differences among structures are seen in the substrate-binding region, as is expected.

3.1. 3-*O*-sulfotransferase

As previously stated, 3-OST is the enzyme that catalyzes one of the final steps in the HS biosynthesis. Different enzyme isoforms transfer the sulfo group from PAPS to the 3-O position of glucosamine units, as can be seen on Figure 5.

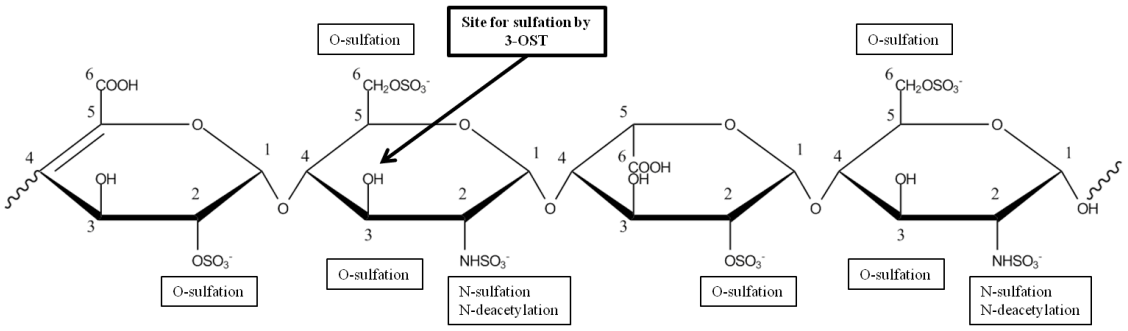

Figure 5 - Representation of the positions in HS which can be modified, with emphasis given to the addition catalyzed by 3-OST.

3-OST is present in seven isoforms with unique expression patterns in human tissues. The different isoforms have a homology of 50-80% in the amino acid sequences of their sulfotransferase domain. 3-OST, as well as the other sulfotransferases involved in the HS biosynthesis, is considered to be a Golgi sulfotransferase.

The different 3-OST isoforms are usually associated with a 3'-phosphoadenosine 5'-phosphate (PAP) molecule, forming a complex. They follow the general structure for sulfotransferases previously described: a spherical shape with a large open cleft across its surface, with an α/β motif in the center of the sphere, consisting of a five-stranded parallel β-sheet surrounded by α-helices. The PSB loop (Lys162-Arg166) is in a strand-loop-helix (SLH3) consisting of the first β-strand and the first α-helix, within the motif. This loop allows for strong hydrogen bonding interactions with the 5'-phosphate of PAPS, which can potentially help position the sulfate group donor (PAPS) in the correct location for catalysis [35, 36].

The binding pocket for the HS substrate is found at the bottom of the open cleft, where the 5'-phosphosulfate of PAPS is found. An α-helix (Ala244-Lys259) is close to the PAPS binding site and is positioned outwards into the open cleft.

The enzyme substrate binds in an extended conformation along the open cleft, which provides access to the sulfate group of PAPS. This fact allows for the sulfotransferase reaction to occur. The 3-OH group of the glucosamine unit (which will be sulfated) is positioned in close proximity to the PAPS-binding site, in the open cleft. In the case of isoform 3, the uronic acid adjacent to the glucosamine unit is the saccharide that establishes more interactions with the protein, which suggests its significant contribution to the substrate specificity [35, 36].

There are a great number of positively charged amino acids around the active cleft, especially when compared with other sulfotransferases, such as the N-sulfotransferase, which is consistent with the higher number of negative sulfo groups in the 3-OST substrate.

Heparan sulfotransferases and cytosolic sulfotransferases are believed to share a common catalytic mechanism, due to the structural similarities and the fact that they all use the same sulfuryl donor, PAPS. This mechanism seems to proceed through a S_N2-like in-line displacement mechanism, where a conserved glutamate residue, present in an exterior 3-stranded antiparallel β-sheet, functions as a catalytic base, deprotonating the 3-OH of the glucosamine unit for nucleophilic attack on the sulfonate group. A bi-pyramidal trigonal transition state (TS) is formed, with the PAP and the acceptor group in the axial positions. A conserved lysine residue on the phosphosulfate-binding loop stabilizes the negative charge on the PAP group, caused by leaving of the sulfate group. This lysine residue forms a hydrogen bond with the bridging oxygen to the sulfur atom. In the specific case of isoform 3, a hydrogen bonding network involving the

glutamate residue which acts as a catalytic base, a histidine and an aspartate is observed. It has been suggested that this hydrogen bond network might help position the glutamate for catalysis or serve as a charge relay system to regulate the pK_a of the glutamate [36]. The active site, as well as relevant interactions established therein, can be seen on Figure 6.

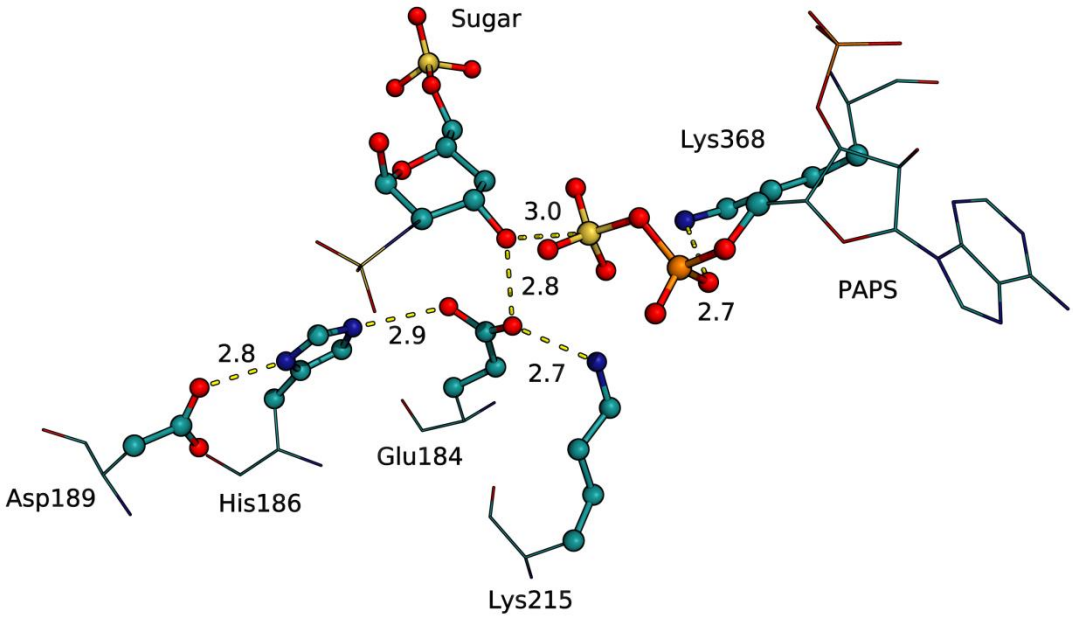

Figure 6 - Three-dimensional representation of 3-O-sulfotransferase active site, taken from the 1T8U structure, displaying the three catalytic residues and the two lysine residues which establish relevant interactions with the reacting fragments, as well as the substrate and co-substrate PAPS. Hydrogen atoms are not represented.

3.2. 2-*O*-sulfotransferase

2-OST catalyzes the transfer of a sulfate group to the 2-OH position of IdoA or GlcA in the HS chain. 2-*O*-sulfated iduronic acid (IdoA2S) is the more abundant and is important in the signal transduction pathways mediated by fibroblast growth factors, since it binds to fibroblast growth factors [37]. This important physiological role is further evidenced by experiments done with animal models, where results have shown this enzyme to have an important role in renal development in mice [38] or axon migration and nervous system development in *C. elegans* [39, 40]. The position modified by 2-OST can be seen on Figure 7.

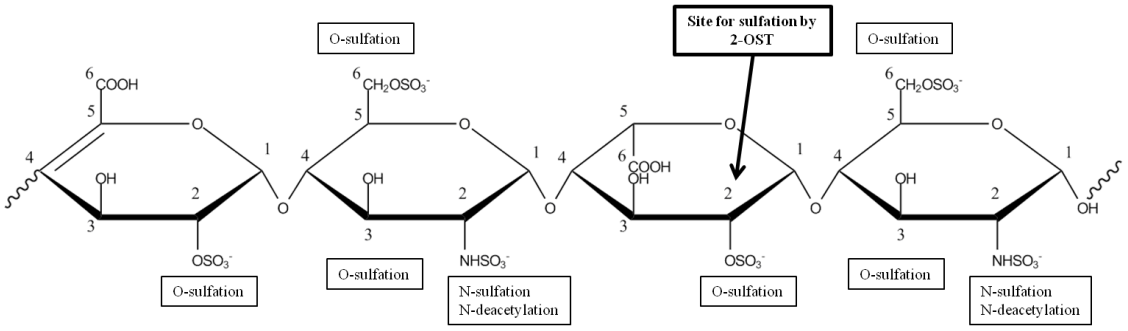

Figure 7 - Representation of the positions in HS which can be modified, with emphasis given to the addition catalyzed by 2-OST.

The enzyme catalytic domain consists of a conical α/β motif, which contains the PAPS-binding loop, similarly to what has been previously described. However, this sulfotransferase presents significant differences, when compared to other sulfotransferases, such as the previously mentioned 3-OST. It presents a low percentage of sequence identity, when compared to 3-OST and the N-sulfotransferase domain of N-deacetylase/N-sulfotransferase-1 (NST). 2-OST also does not contain the structural motif (3-stranded antiparallel β-sheet) which houses the catalytic glutamate base in 3-OST. Interestingly, despite it being a HS sulfotransferase, and thus being a Golgi-associated sulfotransferase, 2-OST shares some structural similarities with cytosolic sulfotransferases, such as conserved residues in the donor and acceptor substrate-binding pockets, as well as the absence of a catalytic glutamate residue, whose function is replaced by a histidine residue, present in 2-OST (His142).

Furthermore, 2-OST seems to function as a trimer, which is significantly different than the other members of the HS sulfotransferase family. The monomers in the trimeric complex seem to position the N termini in close proximity to one another, which seems to indicate that the three molecules would be anchored to the membrane via their N-terminal transmembrane domains. This positioning keeps the three active sites separate, which means that they could potentially function independently [41]. There have been reports that 2-OST forms a complex with C5-epimerase, although it seems unlikely that it would occupy one of the trimeric 2-OST sites [42, 43].

As has been previously mentioned, His142 seems to play a major role in the catalytic reaction of 2-OST. Apart from this residue, Arg80, His140 and Arg288 show potential to participate in the reaction in some way, due partly to the fact that they are located near to the point of the sulfate group transfer. The two arginine residues may be involved in the stabilization of the transition state structure or on the substrate binding, whereas His140 might form a hydrogen bond with Arg80, thus stabilizing its position. The active site, with the interactions mentioned here, can be seen in Figure 8.

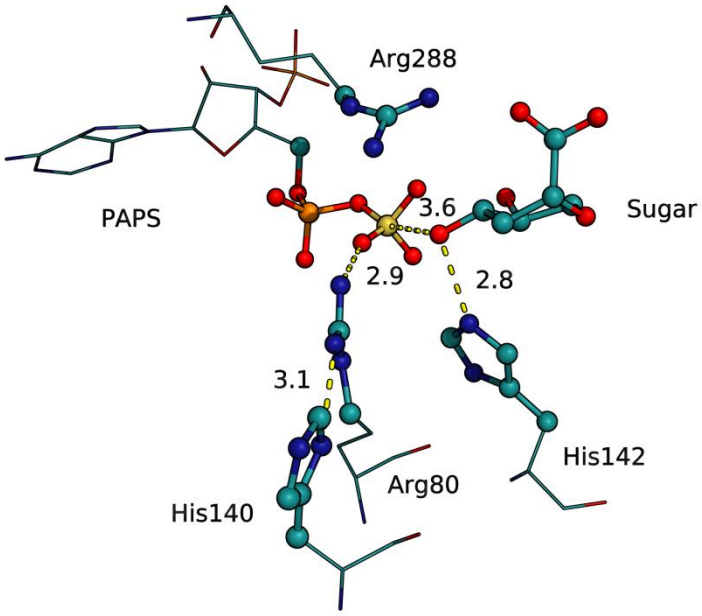

Figure 8 - Three-dimensional representation of 3-O-sulfotransferase active site, taken from the 4NDZ structure, displaying the four catalytic residues, as well as the substrate and co-substrate PAPS. Hydrogen atoms are not represented.

4. Goal

The main purpose of this work is to unravel the catalytic mechanism of the reactions catalyzed by 3-OST isoform 3, which is involved in the 3-*O*-sulfation of a glucosamine unit, and by 2-OST, which is involved in the 2-*O*-sulfation of a glucuronic or iduronic acid adjacent to it. The work also consists of the discovery of compounds with potential 3-OST inhibitor activity, out of a library of compounds.

HSV-1 viral entry requires the interaction of a viral glycoprotein gD with 3-*O*-sulfated HS, synthesized by 3-OST isoform 3. The step prior to this is catalyzed by 2-OST. Therefore, understanding the biosynthesis of 3-*O*-sulfated HS, through the two final steps of its formation, can lead to a new strategy for developing therapeutic agents against HSV-1 infection.

In order to achieve this, the catalytic mechanisms of both enzymes were studied by using a large enzyme model and applying the ONIOM method, with the B3LYP:AMBER methodology. As such, long-distance enzyme-substrate interactions are accounted for, while a great level of detail is maintained at the active site. MD simulations were also

performed, in order to relax the systems in question, as well as, for the case of 3-OST, for determining the protonation states of residues at the active site. Furthermore, pK_a calculations were also performed for both enzymes, in order to assess and confirm the protonation states of key residues. In order to survey a library of compounds for potential 3-OST inhibitors, a virtual screening protocol was applied, in which several compounds were docked into the enzyme active site and ranked according to their binding energy to the enzyme.

II. Computational Methods

With the advent and widespread use of computers, which started in the forties with the invention of the first general-purpose computer [44] and continues on today with the exponential growth in the processing power and decrease in computational cost, came the appropriation of this powerful tool to the benefit of scientific discovery. Among the areas that greatly benefitted from the development of such tool, Chemistry, Physics and Mathematics are some which stand out in taking advantage of it. The combination of these three areas lead to the development of mathematical methods which, allied with the laws of physics, allowed for the study of chemical processes through the use of simulations that explore the virtual environment provided by computers. This allows for the exact modeling of processes which could only, up until then, be explored by using experimental methods, with the computational counterpart being the faster and more cost effective option.

There is, however, a compromise that needs to be made when studying a chemical system with computational methods concerns the fact that processing power is limited. Therefore, a balance needs to be struck between the size of the system and the theoretical model to adopt, taking the available resources into account.

Theoretical models can be based on one of two different theoretical levels: the quantum methods and the classical mechanics methods. The quantum methods encompass the methods based on the wave function, such as *ab initio* or semiempirical methods, and the methods based on the density functional theory (DFT).

The methods based on the wave function have the Schrödinger equation as a starting point [45], which allows to solve the wave function ψ (psi) as a function of the coordinates of the nuclei and electrons, in order to fully characterize the chemical system. The formulation of this equation by Erwin Schrödinger was based on two major theories: the first, established by Max Planck in 1901, dictates that energy varies in discrete quantities (*quanta*), and not in a continuous fashion [46]; the second, first mentioned by Einstein and later completed by Louis de Broglie, states that the aforementioned *quanta* exhibit the properties of both particles and waves (wave-particle duality). However, the Schrödinger equation can only be solved analytically for the hydrogen atom, which means that for more complex polyelectronic systems, simplifications need to be applied to the Schrödinger equation, for it to be solved.

Ab initio methodologies have a significant computational cost, and can therefore only be applied to smaller chemical systems, using these mathematical simplifications and approximations.

Semiempirical methods are also based on the wave function, with similarities with the previously described *ab initio* methodologies. However, some parts of the calculation are approximated, whereas others are replaced by experimental parameters. Considering these simplifications, the computational cost for these methods is greatly reduced. However, this compromise means the introduction of errors in the calculations, which can be reduced by parameterizing and by using a system similar to the one used to parameterize the method. Otherwise, the results obtained through this method would not be viable.

The density functional theory (DFT) [47] has its genesis in the twenties, through the work of Llewellyn Thomas [48] and Enrico Fermi [49]; decades later, in the sixties, this theory was further developed by Pierre Hohenberg and Walter Kohn [50]. In this case, the physical fundamental property is the electronic density. Thus, this method is able to include electronic correlation effects, without resulting in a prohibitive computational cost, as is the case, for example, for *ab initio* methods with a high level of theory, for a significant (50-100) number of atoms. The DFT method can be applied to relatively small systems, consisting of hundreds of atoms.

In the biochemical world, systems are extremely large and other methodologies need to be applied, in order to accurately describe these cases. Larger systems require methods based on Classical Mechanics, such as Molecular Mechanics (MM), which try to predict the structure and properties of said systems based on the laws of classical physics. These methods consider the atom as the smallest unit in the system, representing it by the nuclei properties and the electron distribution. This leads to the inability by these methods to simulate chemical reactions, since these bond breaking and forming events essentially consist in electron transfers. The simulation takes into account the interactions between atoms and, therefore, the mathematical expression used therein includes terms which describe the energy for the lengthening/shortening of bonds, bond angle variation and rotation, as well as terms for van der Waals and electrostatic interactions (non-covalent interactions). This methodology uses constants, in its equations, which are derived from experimental data, as well as data obtained from the aforementioned *ab initio* calculations. Each force field uses specific parameters, as well as equations for calculating the energies mentioned above, with different force fields being specific for specific types of molecules, such as proteins or inorganic molecules, for example.

Methods based on Classical Mechanics have, thus, a very low computational cost, allowing for them to be applied to large systems, such as proteins, allowing for

simulations with a larger timescale. Apart from what has already been mentioned, one of the big downsides of this method is the need for a force field parameterized for the system that is going to be studied, since there is no catch-all force field which would give acceptable results.

All the methods that have been discussed so far are not mutually exclusive. This means that one can combine two or more methods, in order to be able to study a large model representative of the biological environment, with reasonable computational costs, by focusing more accurate methods on the areas of interest, while using less accurate methods for the other areas of the system. The following chapters will expand upon the techniques discussed in this chapter, with special detail to those which will be used throughout this work.

1. Quantum Mechanics

Quantum Mechanics arose due to a need to explain certain phenomena which could not be explained by the reigning theory at the time, Classical Mechanics, which was shown to have severe limitations. Contrary to Classical Mechanics, in Quantum Mechanics, particles are considered to have wave-like as well as particle-like properties. This method allows for a better explanation and description of the atom, particularly when dealing with sub-atomic particles, such as electrons.

1.1. Schrödinger Equation

As has been previously mentioned, the Schrödinger equation can only be solved analytically for the hydrogen atom. In order for its solution to be simplified for more complex systems, several approximations and simplifications must be made. Generally, the equation is time-dependent; however, for stationary systems the time-independent (equation 1) is accurate enough, which makes this approximation valid. Furthermore, relativistic effects are not taken into account using this method, which is only a significant problem for heavy atoms.

$$\hat{\mathrm{H}}\Psi(\mathrm{r}) = \mathrm{E}\Psi(\mathrm{r}) \quad \textbf{(1)}$$

In this equation, E corresponds to the energy in the stationary state, $\hat{\mathrm{H}}$ corresponds to the hamiltonian operator, whereas Ψ is the wave function. The wave function describes the system in question, having as arguments the cartesian coordinates (x,y,z) of all the elements in the system, in a stationary state. This means that one can apply an operator, $\hat{\mathrm{X}}$, and calculate the parameters for the property in question. Ideally, under an infinite number of operations, every single property of a system can be found solely using theoretical calculations, without the use of experimental values.

One of the postulates of Quantum Mechanics, formulated by Max Born, states that by squaring the wave function, the probability density of finding an electron in a given point is obtained (probability density function) [51].

The hamiltonian operator ($\hat{\mathrm{H}}$) is an operator which allows the derivation of the kinetic (particle movement) and potential (Coulomb interactions) energy of a given system, composed of nuclei and electrons. Therefore, in order to calculate the energy of a chemical system, the Schrödinger equation needs to be solved. By decomposing the hamiltonian operator into the different components of the system energy, the resulting operators are the one for the electrons kinetic energy ($\hat{\mathrm{T}}_\mathrm{e}$), the one for the nuclei kinetic energy ($\hat{\mathrm{T}}_\mathrm{n}$), the one for the attraction between nuclei and electrons ($\hat{\mathrm{V}}_\mathrm{ne}$), the one for the repulsion between nuclei ($\hat{\mathrm{V}}_\mathrm{nn}$), and the one for the repulsion between electrons ($\hat{\mathrm{V}}_\mathrm{ee}$), as can be seen on equation 2.

$$\hat{\mathrm{H}} = \hat{\mathrm{T}}_\mathrm{e} + \hat{\mathrm{T}}_\mathrm{n} + \hat{\mathrm{V}}_\mathrm{ne} + \hat{\mathrm{V}}_\mathrm{nn} + \hat{\mathrm{V}}_\mathrm{ee} \quad \textbf{(2)}$$

In order to obtain an energy value close to the ground state of the system, it is necessary to build several different functions; the lower the calculated value, the better the built wave function mimics the real wave function, as suggested by the variational principle, where the calculated $\hat{H}$ value will always be higher than the real energy value for the ground state. These attempts at building a wave function are done by linearly combining a number of mathematical functions, designated by basis functions. The more complete the base function, the better the likelihood of building a wave function that closely resembles the real one.

1.1.1. Born-Oppenheimer Approximation

So far, we have been talking about molecular systems consisting of several atoms; however, calculating wave functions for this type of systems is very hard, due to the fact that the functions in the Schrödinger equation depend explicitly on the coordinates of nuclei and electrons, as well as the correlation between both. In an attempt to counter this issue, the Born-Oppenheimer approximation [52] was developed. Under this approximation, considering that electrons have a much lighter weight when compared to nuclei, and therefore should move at a faster speed, the electron and nuclei movements are considered separately. The approximation considers that the electrons adopt a minimum energy configuration around the nuclei almost instantaneously when the latter move, before they can move significantly. Movement of the two kinds of particles can, therefore, be considered separately, with the electrons moving around fixed nuclei.

In practice, this approximation allows for a simplification of equation 2. The hamiltonian operator can be divided into the electronic hamiltonian, $\hat{\mathrm{H}}_{\mathrm{elec}}$, and nuclear hamiltonian, $\hat{\mathrm{H}}_{\mathrm{nuc}}$, since, for the former, nuclear kinetic energy is not considered, whereas the nuclei repulsion is considered constant. This is shown in equation 3.

$$\begin{aligned} \hat{\mathrm{H}} &= \hat{\mathrm{H}}_{\mathrm{elec}} + \hat{\mathrm{H}}_{\mathrm{nuc}} \\ \hat{\mathrm{H}}_{\mathrm{elec}} &= \hat{\mathrm{T}}_{\mathrm{e}} + \hat{\mathrm{V}}_{\mathrm{ne}} + \hat{\mathrm{V}}_{\mathrm{ee}} \\ \hat{\mathrm{H}}_{\mathrm{nuc}} &= \hat{\mathrm{T}}_{\mathrm{n}} + \hat{\mathrm{V}}_{\mathrm{nn}} \end{aligned} \qquad \textbf{(3)}$$

When taking into account a system with fixed nuclear geometry, the hamiltonian is represented by equation 4.

$$\begin{aligned} &\hat{\mathrm{H}} = \hat{\mathrm{H}}_{\mathrm{elec}} + \hat{\mathrm{V}}_{\mathrm{nn}}, \text{ with} \\ &\mathrm{E}_{\mathrm{tot}} = \mathrm{E}_{\mathrm{elec}} + \hat{\mathrm{V}}_{\mathrm{nn}} \end{aligned} \qquad \textbf{(4)}$$

The Born-Oppenheimer approximation makes it possible to solve the Schrödinger equation for systems consisting of more than one nucleus but only one electron, such as, for example, the dihydrogen cation.

1.1.2. Nuclear and electronic hamiltonians

Due to the presence of the repulsion between electrons, electronic energy is difficult to calculate. Therefore, there are methods based on the wave function or on the density functional which try to calculate it. In these cases, the total energy corresponds to a system with fixed nuclear geometry, as represented by equation 4. However, in order for the system to be fully characterized, it has to include the nuclear hamiltonian (equation 3), which includes the terms in the nuclear hamiltonian, the nuclear kinetic energy ($\hat{T}_n$) and the repulsion between nuclei ($\hat{V}_{nn}$). This is done by using a nuclear Schrödinger equation, which can be decomposed in translation, rotation and vibrational components. In the stationary state, the nuclear energy is only dependent on the vibrational component, since the translational and rotational components are not present. Therefore, in order to obtain an accurate depiction of the system, one should take into account the vibrational effects in the system, through calculation of vibrational frequencies.

1.2. Wave function-based methods

As has been previously mentioned, the Born-Oppenheimer approximation is still not enough to solve the Schrödinger equation for polyelectronic systems, since it only allows for solving the equation for one electron systems with more than one nucleus. The difficulty in solving the electronic hamiltonian operator still remains, due to the term referring to the repulsions between electrons.

In order to attempt to solve this problem, Hartree [53] initially suggested that the electronic correlation could be neglected, causing each electron to interact with the field created by the average distribution of all other electrons, and calculating the wave function as a product of the monoelectronic wave functions for each electron. This would only take into account the monoelectronic component of the system, with the bielectronic component corresponding to the electronic correlations not being considered.

Even though this approximation allowed for solving the Schrödinger equation, it could not be considered as a viable methodology, due to the fact that it did not consider the electronic correlation effect. Therefore, Hartree later suggested that those effects could

be taken into account by considering that each electron interacts with an average field, which is created by the nucleus attraction and the repulsion of the average distribution of all other electrons. Although this method still presented severe limitations, such as ignoring the spin variable, it was an important stepping stone in developing more robust methods, such as the iterative self consistent field (SCF) method.

The Hartree-Fock method came as a follow up of the initial Hartree method. This was possible due to the introduction of the Slater determinant [54]. Whereas the Hartree method did not respect the principle of antisymmetry of the wave function, this determinant of one-particle orbitals did, which meant that it was suitable for applying the variational principle. Therefore, all the wave functions used have to obey the Pauli exclusion principle. This method has, in its basis, the fundamental principles introduced by Hartree, such as the treatment of electronic correlation as an interaction between electrons (in spin orbitals) and the average potential resulting from the nuclei attraction and the average repulsion with other electrons.

The Hartree-Fock equation is a reformulation of the electronic hamiltonian, distributing the electrons in spin orbitals. For each electron:

$$\hat{F}\phi_i = \varepsilon_i \phi_i \qquad \textbf{(5)}$$

ϕ_i corresponds to the wave function for the spin orbital of electron i (Hartree-Fock molecular orbitals), whereas ε_i is the energy corresponding to the spin orbital. $\hat{F}$ is the Fock operator generated by the spin orbital and can be represented as such:

$$\hat{F}_i = \widehat{h_i} + \sum_{j=1}^{n} [\, 2\,\hat{J}_j(i) - \widehat{K_j}(i)] \qquad \textbf{(6)}$$

The term $\widehat{h_i}$ corresponds to the monoelectronic hamiltonian for electron i, n is the total number of occupied orbitals, whereas $\hat{J}_j(i)$ is the Coulomb integral, defining the electron-electron repulsion energy and $\widehat{K_j}(i)$ is the exchange operator, which defines the electron exchange energy, representing the energy resulting from the exchange between electrons in two different orbitals.

1.2.1. Variational Principle and the Self Consistent Field (SCF) method

The variational principle, which has been previously briefly touched upon, serves as a basis for most of the methods which try to approximate a solution to the Schrödinger equation. According to this principle, the energy obtained for a system, calculated by using a tentative wave function, will always be the same as or higher than the energy of the system obtained by using the real wave function. Therefore, if the wave function is altered until the energy value is at a minimum, the obtained energy will be close to the real one.

The Self Consistent Field (SCF) method is based on the variational principle and is used to calculate the energy of a system, using the Hartree-Fock equations. Since the Fock operators are not independent between them, the energy could only be calculated if the values for the wave functions of all electrons were known. The SCF method iteratively creates tentative functions by choosing a number of spin orbitals, relevant to the system. Therefore, the Fock operators for each electron are calculated based on this estimative and are included in the Fock matrix, which allows for the solution of the Hartree-Fock equation. New and lower energy spin orbitals are obtained through this solution and the process is repeated using these, successively, until two successive iterations have a difference between each other lower than a defined threshold.

1.2.2. Hartree-Fock-Roothaan-Hall method

Although the SCF method was very well received, due to the precision of its results, the Hartree-Fock method turned out to only be analytically solvable for single atoms, due to the spherical symmetry of atomic orbitals.

A new iteration of the Hartree-Fock method was proposed by Roothaan and Hall, in which the linear combination of atomic orbitals would be applied to this method [55, 56]. The atomic orbitals are similar to the basis functions, the set of functions which apply to a single electron, centered in the atomic nucleus. The accuracy of the results is highly dependent on the number of basis functions, since the more complete the base function, the closer to the real wave function the calculated wave function will be. However, this increase in accuracy comes with a very significant cost, since the computational cost of using these methods increases greatly with the number of basis

functions, in already costly *ab initio* methods. Therefore, a compromise must be made between the accuracy of the calculations and the available computational power.

1.2.3. Semi-empirical methods

Semi-empirical methods are quantum models based on the wave function, namely on the Hartree-Fock theory. However, contrary to the other methods discussed previously, these make several approximations in the calculations, which means that, in order to be comparable to experimental data or results obtained through *ab initio* methods, they include adjustable parameters, obtained via parameterization. These approximations and omissions include the zero differential overlap (ZDO) [57], which ignores integrals concerning the products of different basis functions which depend on the same electronic coordinates. This leads to the exclusion of two-electron repulsion integrals, which are replaced by empirical parameters.

Since this work is not going to use this type of methods, they will not be further expanded upon.

1.3. Density Functional Theory (DFT)

As has been repeatedly stressed throughout this work, the results obtained by the wave function-based methods are extremely accurate, including electronic correlation effects. However, the downside to this accuracy is the very high, sometimes prohibitive, computational cost.

The density functional theory (DFT) [47] aims to lower the complexity associated to the poly-electronic wave function, by calculating the energy through electronic density ($\rho(r)$). The upside to this is the dependency of this parameter of only 3 coordinates (X,Y,Z), independent of the number of electrons, when compared to the previously discussed wave function-based methods, which depends on 4n coordinates, including spin, where n is the number of electrons in the system. DFT has a constant number of variables, regardless of the number of particles in the system.

By using the electronic density to calculate the energy, Hohenberg and Kohn proposed that it is possible to determine the static external potential, which is due to the coulomb attraction between electrons and nuclei. It is possible to infer, then, that the electronic density determines the charge and position of all the nuclei. Therefore, by knowing the position and charge of all the nuclei and the number of electrons in the system, it is possible to determine the system hamiltonian and, thus, the wave function and all its associated properties, such as energy, which means that, in this case, the energy and all the other properties are electronic density functionals.

Hohenberg and Kohn postulated that any property of a system in the ground state can be calculated from the electronic density in this state; furthermore, they demonstrated that by minimizing the energy in the ground state, one can accurately determine the electronic density of a non-degenerate ground state in the absence of a magnetic field. Ergo, if the dependency of energy as a function of the electronic density is known, through energy minimizations, the ground state electronic density can be obtained. However, although this process seems relatively straightforward, this is where the DFT method somewhat fails, since the functional is not entirely known nor is it known how to obtain improved density functionals.

1.3.1. Kohn-Sham method

In order to try and solve the issue that pertains to the lack of knowledge of the functional which relates energy and the electronic density function, Walter Kohn and Lu Sham proposed a simplification [58]. This consists of considering a hypothetical chemical system, with non-interacting electrons, which were subject to an external potential, so that the density of this system was equal to the one in the ground state of the chemical system that one intends on studying, in this case interacting electrons.

The lack of interaction between electrons in the hypothetical system makes it possible to find a Slater determinant consisting of a number of spin orbitals equal to the number of electrons in the system. This Slater determinant, due to the fact that both the systems have the same electronic density, is the exact wave function for the real system in the ground state, and, through the spin orbitals, the Kohn-Sham orbitals (mathematical functions which reflect the ground state electronic density) can be obtained. The biggest challenge in this method is finding a set of Kohn-Sham orbitals which result in an electronic density similar or close to the real system.

As was previously stated, the Born-Oppenheimer approximation considers that the nucleus-nucleus repulsion is constant. The consideration of the hypothetical system, wherein electrons do not interact with each other, allows for the separation of the terms in the calculation, as shown in equation 7. One of the parts of the equation concerns a system of non-interacting electrons, allowing for it to be solved in an exact fashion, since its hamiltonian is a result of the sum of mono-electronic operators, whereas the other consists of a correction term. The exchange-correlation energy ($\mathrm{E_{xc}}$) term attempts to compensate the error introduced by the difference between the calculated energy and the real system, containing electrons which interact amongst themselves. This difference manifests itself through the difference between the kinetic energy (real vs hypothetical system), as well as the coulomb repulsion resulting from the interaction between electrons. However, the precise description of the exchange-correlation energy ($\mathrm{E_{XC}}[\rho]$) is not known, which means that methods based on DFT have to use an approximation description. Equation 7 represents the energy of the system, calculated by this method.

$$\mathrm{E}[\rho] = \mathrm{T_s}[\rho] + \mathrm{V_N}[\rho] + \mathrm{J_{ee}}[\rho] + \mathrm{E_{XC}}[\rho] \qquad \textbf{(7)}$$

The $\mathrm{T_s}[\rho]$ term corresponds to the non-interacting electrons kinetic energy, $\mathrm{V_N}[\rho]$ relates to the energy of the nucleus-electron interaction, $\mathrm{J_{ee}}[\rho]$ corresponds to the coulomb electronic repulsion; the term $\mathrm{E_{XC}}[\rho]$, as mentioned before, corresponds to the exchange-correlation energy and can be separated into $\mathrm{E_C}[\rho]$ and $\mathrm{E_X}[\rho]$, the correlation and exchange functionals, respectively.

In order to build the Kohn-Sham orbitals which accurately describe the system by having a density close to the real system, the variational principle must be, once again, applied. However, this method presents a slight difference, since $E_{XC}[\rho]$ can adopt a lower energy than the real system. The orbitals should correspond to a density which minimizes the system energy and should be obtained by iterating the self-consistent field, in an analogue way to what was described for Hartree-Fock methods. Similarly, and starting on an initial electronic density, the Kohn-Shan orbitals for that density are calculated; those orbitals are used in order to calculate an electronic density closer to the real system, repeating the process until the density and the exchange-correlation energy variations reach a lower threshold. After this iterative step, the electronic energy is calculated.

1.3.2. Approximated Functionals

As was previously mentioned, the main problem of the DFT methods is the fact that the exact value of $E_{XC}[\rho]$ is not known; thus, the methods have to use an approximate description of the term. Therefore, in order to calculate the properties of the system through this method, corrections have to be introduced, in order for the term to be accurately described. As was previously mentioned, and can be seen in equation 8, $E_{XC}[\rho]$ can be decomposed into the sum of two independent functionals, the exchange functional ($E_X[\rho]$), concerning interactions between electrons with the same spin, and the correlation functional ($E_C[\rho]$), concerning interactions between every electron:

$$E_{xc}[\rho] = E_x[\rho] + E_c[\rho] \quad \textbf{(8)}$$

Different types of approximations can be made for each of the components:

- Local (Spin) Density Approximation (L(S)DA) which explicitly depends on density or spin density;

- Generalized Gradient Approximation (GGA), which depends on electronic density and its gradient;

- Meta-GGA, the more recent approximation based on GGA, including the second derivative of the electron density instead of only the density and the first derivative in the exchange-correlation potential, as is the case with GGA.

The former (LDA/LSDA) consider the density as a homogenous electron gas (HEG). Therefore, the electronic density is taken as a functional that varies very slowly. There is no analytical solution for the correlation functional and, therefore, in order to use this type of functionals, there are several precise energy values for different density HEG which need to be used. Although LDA/LSDAs show adequate results for systems where the electronic density is constant, such as metals, this type of functionals is not adequate for more complex systems, where the density varies as a function of the coordinates.

In order to overcome the issue with this type of functionals, others were developed, allowing for the correlation and exchange energies to depend on not only the electronic density, but the gradient as well. There are several GGA correlation functionals and

several exchange GGA functionals, which can be combined amongst themselves. This is the case for several widely used in computational chemistry, such as the combination of the gradient correlation functional LYP (Lee, Yang and Parr) with the gradient exchange functional by Becke, leading to the BLYP functional.

There are also hybrid functionals. These are the result of the combination of the exchange-correlation functionals with an exact exchange portion from Hartree-Fock. The exact exchange functional is expressed in terms of the Kohn-Sham orbitals, rather than the density. An example of this is one of the most ubiquitously used functionals, B3LYP, which combines the previously mentioned LYP correlation functional with the three parameter Becke functional (B3). The equation for this functional can be seen in equation 9.

$$E_{XC}^{B3LYP}[\rho] = (1-\alpha)E_X^{LSDA}[\rho] + aE_X^{HF}[\rho] + bE_X^{B}[\rho] + cE_C^{LYP}[\rho] + (1-c)E_C^{VWN3}[\rho] \quad \textbf{(9)}$$

VWN3 is the local-density approximation to the correlation functional by Vosco, Wilk and Nusair [59]; a, b and c are empirical parameters (respectively, 0.20, 0.72 and 0.81), determined experimentally.

1.4. Basis functions

Basis functions, as has been previously mentioned, are a set of known functions which try to describe the behavior of an unknown function, such as atomic, molecular or Kohn-Sham orbitals, and are necessary in order to use *ab initio* or DFT methodologies. The more complete or higher the number of basis functions, the better the description of the system; however, this better description comes at a steep computational time cost. Therefore, the choice of the base function has to take into account the weight of these pros and cons.

The two most relevant types of basis functions are the Slater-type orbitals (STO) and the Gaussian-type orbitals (GTO).

STO functions vary exponentially with the electron-nucleus distance and provide a better description for atomic orbitals, due to their similarities with one-electron orbitals. However, for atomic orbitals with more than two centers, they are not adequate since the electronic repulsion integrals are difficult to calculate.

Unlike STOs, the bi-electronic integrals can be analytically calculated using GTOs. However, these functions are not able to adequately describe electronic density, particularly very close and very far from the nucleus, which means that the basis set needs to be larger. Nonetheless, even with these limitations, GTOs are preferred to STOs, since their mathematical simplicity outweighs the cost of using a larger basis set, thus reducing computational cost.

By linearly combining a set of GTOs (primitives), in which the coefficients are fixed, in order to form a smaller base function, the GTOs can be improved and better describe the system in question. The resulting basis functions are referred to as contracted GTOs (CGTOs). This combination leads to a better description of orbitals with electrons close to the nucleus.

In order to accurately depict molecules, basis functions need to allow for the flexibility in orbitals, so that the chemical bond between the atoms is adequately described. This flexibility can be increased through the use of more basis functions for each orbital and through the inclusion of polarization functions and diffuse functions.

The smaller basis functions are single-ζ functions (SZ), which only have one base function for each occupied orbital. In the case of single-ζ CGTO functions, since they are meant to mimic STOs, these are represented by STO-nG, where n is the number of primitive GTOs used to build the CGTO. So, as an example, STO-3G would be a SZ function where each atomic orbital has one CGTO, which results in the contraction of 3 GTOs.

By using two (double-ζ) or three (triple-ζ) basis functions for each orbital, it is possible to obtain more flexibility and a better electronic distribution. However, since the valence orbitals are the ones with more chemical relevance, double- and triple-ζ basis functions are only used for these.

Generally, the nomenclature of basis functions for valence orbitals follows the following syntax, proposed by Pople [60]:

x-yzG

Where x represents the number of primitive Gaussians which composes each core atomic orbital base function and y and z describe the number of basis functions which compose each valence orbital. The G indicates that the orbitals are of the Gaussian type. The length of the name after the dash (in the example above, "yz") indicates if the

base function is a double-ζ (two numbers after the dash, y and z), triple-ζ (three numbers after the dash), and so on.

The nomenclature also takes into account the use of polarized functions, through the use of parenthesis. For example, taking the generic example above, if the d orbitals were to be included, the nomenclature would be x-yzG(d). The use of diffuse functions is indicated by one or two plus signs ("+" or "++") before the G, depending if the functions are used, respectively, in heavy atoms or hydrogen atoms.

2. Hybrid Methods

Enzymes are complex biomolecules, usually consisting of thousands of atoms, which catalyze a broad spectrum of reactions. Their large sizes make it impossible to simulate them using, exclusively, Quantum Mechanics (QM) methods based on the Born-Oppenheimer approximation. However, in order to accurately study the breaking and forming of bonds in their active centers, it is necessary to use methods that allow for the redistribution of electrons among atoms as a reaction proceeds.

Hybrid methods, generally combining QM and MM (QM/MM methods) have the possibility to combine two or more computational techniques, which allows the study of very large systems with high accuracy in their active sites. In these methods, the system is divided into an electronically important region that requires a quantum chemical treatment, the Quantum Mechanics (QM) region, and a surrounding one that requires a classical mechanical treatment, the Molecular Mechanics (MM) region. In general, the QM region includes all the atoms that are directly involved in the chemical reaction under study, while the MM region contains all the other atoms of the system, where the formation and breaking of bonds do not occur. In the hybrid QM/MM method, therefore, the enzymatic reaction is treated as a transformation involving only the QM region, consisting of at least partially the substrate, residues direct or indirectly involved in the reaction catalyzed by the enzyme and sometimes a cofactor. The MM region includes the surrounding enzymatic environment and sometimes some solvent molecules, which may influence the QM region by the establishment of long-range interactions [61].

The coupling between the two regions of the system is an important issue. It must be capable of accounting for the bonded and non-bonded interactions. The simplest

approach to treat the covalent interaction between the two regions is to use link atoms, as shown on Figure 9 [62]. These are usually hydrogen atoms but can be any atom that mimics the part of the system it substitutes. This, however, may introduce polarization effects that destabilize the system. Another way to characterize the interactions of the atoms in the boundary between the two regions is by using the frozen orbital approximation, in which localized, frozen molecular orbitals, built by linear combination of two hybrid orbitals located in each of the atoms of the covalent bond, are used as the interface between the QM and MM region. This method is rigorous but hard to implement; furthermore, the parameters obtained for the covalent bond may not accurately describe the bond in the system being studied.

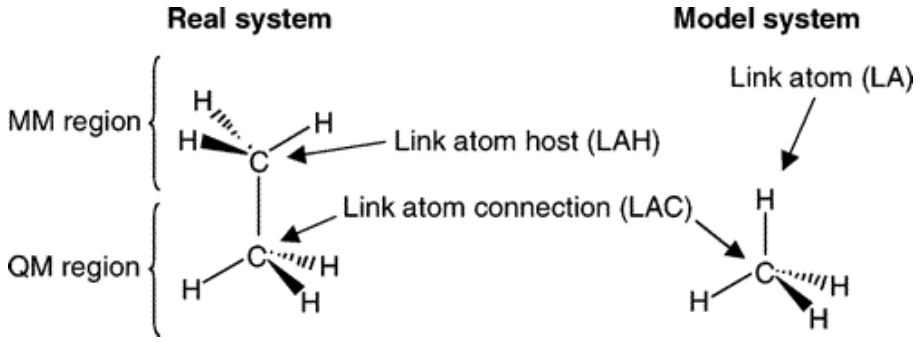

Figure 9 - The nomenclature used in the QM/MM methods is displayed on an ethane molecule, as an example. Two regions, one using QM methods and another using MM methods, are defined and a link atom is used in the boundary, in the model system. Adapted from **[63].**

The electrostatic interaction between the two layers can be treated in different ways. The simplest way is to describe the interaction between the QM and MM regions only at the MM level. The electrostatic interaction is evaluated as the interaction of the MM partial charges with partial point charges assigned to the atoms in the QM region. This approach is called classical or Mechanical Embedding (ME). Another approach involves the interaction of the charge distribution of the MM region with the actualized charge distribution of the QM region. The partial charges from the MM region, in this case, are included in the QM Hamiltonian. The electrostatic interaction is, therefore, more accurately described and the wave function is able to respond to the charge distribution of the MM region. This approach is called Electronic Embedding (EE) [64].

The QM/MM schemes can be additive or subtractive. In the first case, the energy for the QM and the MM region are computed and added to a coupling term, in order to obtain the total energy of the system, as is represented by the hamiltonian in equation 10.

$$\hat{H}_{QM/MM} = \hat{H}^{L}_{MM} + \hat{H}^{S}_{QM} + \hat{H}^{coupling}_{MM} \qquad \textbf{(10)}$$

L represents the large model, treated with MM, whereas S represents the small model, treated with QM. The coupling term is difficult to accurately calculate, especially when in the presence of link atoms or when electrostatic perturbations are included in the QM Hamiltonian.

In the subtractive schemes case, the energy of the entire region is computed in MM. The energy of the inner layer is calculated both at the QM and MM level. The inner layer energy calculated through MM is subtracted from the other two energies calculated in the case of two-layer systems, as shown in equation 11.

$$\mathrm{E}_{\mathrm{QM/MM}} = \mathrm{E}^{\mathrm{L}}_{\mathrm{MM}} - \mathrm{E}^{\mathrm{S}}_{\mathrm{MM}} + \mathrm{E}^{\mathrm{S}}_{\mathrm{QM}} \qquad \textbf{(11)}$$

This scheme is the one used in several methods, such as the integrated molecular orbitals (MO) + MM method (IMOMM) [65], the integrated MO + MO method (IMOMO) [66] and the "our own n-layered integrated MO and MM" method (ONIOM) [67], which was used in this work.

This type of method does not need a parameterized expression describing the interaction of the different regions. However, the fact that all the atoms in the inner layer need to be described at the MM level can pose a problem, especially when dealing with non-typical atom types such as metals. To solve these specific cases, it is necessary to add external parameters for these non-typical atoms. It is also important to take particular attention when the influence of the MM layer on the electronic structure of the QM layer is significant, since the electrostatics on the MM region cannot polarize the wave function in the QM region.

3. Docking and Virtual Screening

In protein-ligand docking, which is the type of docking that will be used throughout this work, different conformations of the chosen ligand are generated and are placed in a chosen area of the fixed protein. A molecular docking process requires a search algorithm and a scoring function, in order to try and find the binding pose of the ligand and its binding energy [68]. The docking process consists of the prediction of the preferred pose and conformation of the ligand, in relation to a receptor (an enzyme, in this case), when the binding of both forms a stable complex. Docking protocols can be performed keeping both receptor and ligand flexible or with only one of the components

flexible. The method which implies less complexity is the one with only the ligand (which is smaller) flexible and, as such, is the one which should be used in virtual screening campaigns, which will be explored ahead.

The different conformations of the ligand are generated by the search algorithm. There are three different types of search algorithms:

- The systematic algorithms, which try to explore every degree of freedom in a molecule. They can be classified into: conformational search methods (brute force, rotating every bond in every possible way in order to find binding poses), fragmentation (incremental docking of fragments of the ligand onto the active site) and database methods (libraries of known conformations in order to reduce the flexibility of the ligand);

- The stochastic algorithms, which perform random changes to the ligand which is going to be docked, either through Monte Carlo Methods (based on a Boltzmann probability function), Genetic Algorithm methods (which apply ideas derived from genetics into docking) or Tabu Search methods (imposes restrictions based on already explored conformations, in order to analyze new regions).

The scoring function ranks the ligand conformations, enabling the distinction between true binding modes and binding modes which are not relevant for the docking. Scoring functions include some approximations and simplifications, otherwise the computational cost would be too large. Scoring functions are generally divided into four major classes:

- Force Field-Based Scoring, which quantifies the interaction energy between receptor and ligand and the internal energy of the ligand, through a combination of a van der Waals and electrostatic energy terms. The functions and parameters are comparable to the ones used in molecular mechanics, albeit with some additional terms;

- Empirical Scoring Functions, which predict the binding energy based on a number of terms (such as hydrogen bonds) related to experimental data;

- Knowledge-Based Scoring Functions attempt to reproduce structures determined experimentally, through the use of simple atomic interactions-pair potentials;

- Consensus Scoring, which is essentially a combination of the other three.

Autodock, one of the more ubiquitous docking programs, uses a Genetic Algorithm (GA) to generate conformations, ranking and scoring them through a Force Field-Based scoring function, as shown in equation 12.

$$\Delta G = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{L-P} - V_{unbound}^{L-P} + \Delta S_{conf}\right) \quad \textbf{(12)}$$

V corresponds to the pair-wise interactions, whereas ΔS_{conf} is a factor concerning the loss of conformational entropy upon binding. L and P correspond to the ligand and protein, respectively.

Each of the pair-wise interaction terms includes terms for dispersion/repulsion, hydrogen bonding, electrostatics and desolvation, as shown in equation 13.

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{6}}\right) + W_{elec}\sum_{i,j}\frac{q_i q_j}{e(r_{ij})r_{ij}} + \sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{-r_{ij}^2/2\sigma^2} \quad \textbf{(13)}$$

W corresponds to weighting constants, optimized to calibrate the empirical free energy, based on experimentally determined binding constants. The first term corresponds to van der Waals dispersion/repulsion interactions, with parameters based on the Amber force field. The second term is an H-bond term, with a correction ($E(t)$) for a deviation (t) from the ideal H-bonding geometry. The third term is a Coulomb potential for electrostatics, whereas the final one is a desolvation potential based on the volume (V) of atoms which surrounds a given atom and shelter it from the solvent (S); in this term, σ corresponds to a distance-weighting factor.

Virtual screening applies the docking methods discussed so far to a large number of compounds, screening and ranking them based on their free binding energy to a chosen region of the protein in study. It is, in essence, a repeated application of the protein-ligand docking protocol to a library of compounds, which, due to the usually large number of compounds, has to rely on the less computationally demanding options for docking.

This method is widely used in drug discovery studies because it allows for an initial screening and filtering of compounds, making the following steps in the drug design process easier and more effective. It reduces the number of compounds being studied by more complex experimental techniques.

III. Procedure

The following chapter will provide an outline to the work that was performed throughout this master thesis, detailing all the methods used. There were two main parts to this work, which involved the study of the two enzymes that have already been previously mentioned, 3-OST (isoform 3) and 2-OST.

1. 3-*O*-sulfotransferase

The study of the catalytic mechanism of 3-OST is a very important step in the development of potential treatment for infection by HSV-1, since this enzyme has very significant potential as a therapeutic target and provides valuable information to the catalytic mechanism of the sulfotransferase family. In order to do so, it is first necessary to create a model which accurately mimics the enzyme biological environment. After choosing the PDB structure correspondent to the enzyme, it is necessary to investigate which residues are relevant for catalysis, which has been briefly touched upon in the introductory chapter. Furthermore, since, in the specific case of this enzyme, the important residues can adopt several different protonation states, it is important to assess which protonation states are more relevant for catalysis. This was made possible through the use of web-based pK_a prediction tools, such as H++ [69-72], which was used throughout this work, as well as through the use of Molecular Dynamics (MD) that clarified the results obtained through the H++ tool. Hydrogens were added to the model based on the protonation states of the residues, using the xleap program from the AmberTools package. In order to perform the MD simulations, the PAPS and disaccharide had to be parameterized using Hartree-Fock methods, using the antechamber program, part of the Amber suite [73]. Furthermore, the MD simulations also had the purpose of relaxing the system, since the crystallographic structure may not entirely reflect the biological environment, especially considering the fact that small changes had to be made to the structure, such as the deletion or modeling of portions of the substrate molecules. Several structures were also extracted from the MD simulations, in order to later study the dependency of the energetics of the mechanism on the conformation used.

After these first steps, the model for the aforementioned hybrid QM/MM method was built, taking into account the key residues, which directly or indirectly participate in the catalytic reaction and have to, therefore, be included in the high layer, treated with the DFT method, using the B3LYP functional. The rest of the enzyme was kept in the low

layer and treated with molecular mechanics. The model was then optimized with the 6-31G(d) base function, starting from a equilibrated structure extracted from the MD simulations. Using the optimized structure, a potential energy surface (PES) scan was run, ranging from reactants to products, in order to study the catalytic mechanism.

After clarification of the catalytic mechanism, a virtual screening protocol was run, using a VMD [74] plugin developed in house (VsLab [75]) which consists of an implementation of Autodock in order to perform high-throughput screening of compounds. The library of compounds used was a natural compound library put together in house and a collection of compounds with known sulfotransferase inhibitor activity.

1.1. Building the model

The model used for this study was obtained from the 3-OST isoform 3 X-ray crystallographic structure with the Protein Data Bank (PDB) [76] ID 1T8U. There are, up until the time of writing (September 2015), two structures corresponding to the human 3-OST isoform 3 (PDB IDs: 1T8T and 1T8U) at 1.95 Å resolution, which only differ amongst themselves in the presence (or absence) of the substrate molecule, with the former not having the substrate molecule. Both the structures are native, not containing any mutations. Therefore, the latter was used, since it already has the modeled HS tetrasaccharide, which will be sulfated during the catalytic reaction, and the PAP molecule bound. This HS behaves as two separate disaccharide units, one on the reducing end consisting of iduronic acid-glucosamine and one on the non-reducing end consisting of glucosamine-uronic acid. Uronic acid establishes the most extensive interactions with the protein and the glucosamine unit is the one that will be sulfated. As the disaccharide unit on the reducing end presents fewer hydrogen bonding donors and acceptors, and has only a spectator role during the catalysis, it was not considered in the model used in the calculations. The glucosamine-uronic acid disaccharide was used as substrate in the study of the catalytic mechanism of the enzyme. A sulfate group was also modeled onto PAP in order to form the sulfate donor PAPS. Since this group is small (4 atoms), the modeling tool contained within the GaussView program from the Gaussian 09 [77] software was used. The group was added to the oxygen atom closest to the position that would be sulfated in the catalytic reaction. The chain B from this crystallographic structure was used. Although the crystal behaves as a dimer, the

contact interface between the two chains present in the crystallographic structure is located far from the active site, which suggests that the activity of each active site center is independent to the presence of the other chain. Chain B was chosen due to the fact that it already has the substrate molecule modeled.

Based on previous experimental mutational and structural analysis, as well as comparison with other sulfotransferases, the key residues for the catalytic mechanism were chosen, either because they are expected to participate directly in the reaction or because they stabilize the high negative charge of the active site, which is due to the substrate sugar and PAPS molecules: Lys162, Glu184, His186, Asp189, Lys215 and Lys368. All but the lysine residues are expected to participate, directly or indirectly, in the reaction, whereas the lysine residues are expected to stabilize the negative charge.

1.2. pK_a estimation

The pK_a for the residues which directly participate in the catalytic reaction were estimated using the previously mentioned web-based server H++ (Salinity = 0.15, Internal Dielectric = 10, External Dielectric = 80 and pH = 6.5). [70-72]

In order to obtain a clearer picture of the protonation states of the relevant residues (Glu184, His186 and Asp189), Molecular Dynamics (MD) simulations were performed on the system. Three models were built: one with neutral Glu184 and His186 and negatively charged Asp189 (A); one with neutral His186 and negatively charged Glu184 and Asp189 (B) and one with positively charged His186 and negatively charged Glu184 and Asp189 (C). The different protonation states used for this work are detailed in Figure 10. Other options for the protonation states were discarded as they would not be adequate for the catalytic reaction to take place. Hydrogen atoms were added using the xleap program from the AmberTools package [73].

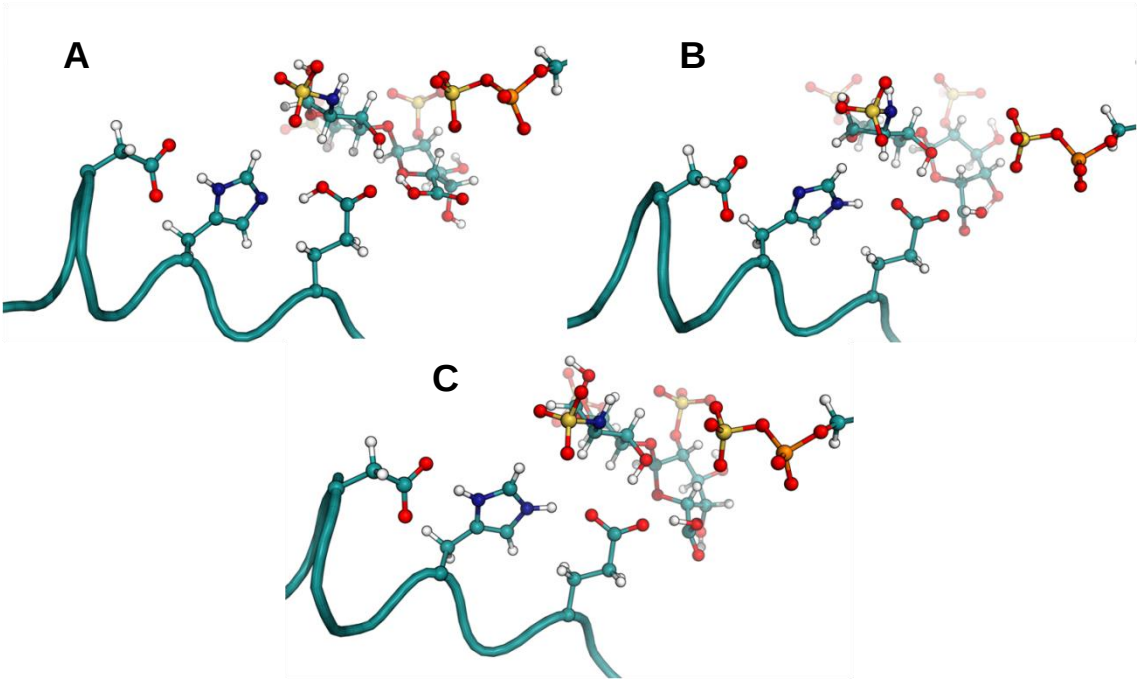

Figure 10 - Representation of the different catalytic triad protonation states studied in the MD simulations. In Figure 10A, the Glu184 and His186 residues have a neutral charge, while the Asp189 has a negative charge. In Figure 10B, both the Glu184 and Asp189 have a negative charge, while the His186 is neutral. In Figure 10C, both the Glu184 and Asp189 have a negative charge, while the His186 has a positive charge. The three Figures also show part of the substrate and of PAPS.

Two force fields were used: the AMBER 2003 force field (parm03) [73] for proteins and the general AMBER force field (gaff) [78] for the substrate and PAPS, which were parameterized using the Antechamber program [79]. Explicit solvent, consisting of TIP3P water molecules in a truncated rectangular box with a minimum 12 Å around the protein, was used and chloride counter-ions were added in order to neutralize the charge of the system, leaving the final model atom count at 51881 atoms. Atomic charges were calculated using a QM optimization calculation at the HF/6-31G(d) level, followed by a RESP (Restrained Electrostatic Potential) calculation. [80].

Two separate minimization stages were performed, first keeping the protein fixed and minimizing the water and counter-ions and then, in the second step, minimizing the whole system. The minimized structure was subsequently used to run a MD simulation for 200 ps using the NVT ensemble and periodic boundary conditions, where the temperature was increased from 0 K to 310.15 K. This equilibration stage preceded the second stage, where the NPT ensemble was used, with 10 ns duration. The temperature was kept constant throughout the MD simulations by use of a Langevin thermostat [81], with a collision frequency of 1 ps^{-1}. The Particle-Mesh Ewald (PME) method [82] was used to treat long-range interactions, and the non-bonded interactions were accounted for in a 10 Å radius. The SHAKE algorithm was used in order to constrain the bond lengths involving hydrogen atoms [83]. The time step used was 2 fs.

The simulations were run using the sander module and analyzed using the ptraj tool, both in the AMBER 9 package [73].

At the first stage, the system is at a constant volume while being heated. This keeps certain elements from moving apart, while preparing the system for the simulation of the biological phenomena, which happens at the second stage. In the production phase, temperature and pressure are kept constant, but the volume is not. This allows for a realistic representation of the solvent molecules density and, therefore, a more accurate description of the biological system being studied.

The results of the MD simulations were analyzed using the ptraj tool, which is included in the Amber package [73] that analyzes the coordinates and trajectories obtained from the simulations.

1.3. QM/MM model and calculations

The QM/MM calculations were run using the Gaussian 09 software [77]. The initial structure was obtained from the equilibrated aforementioned MD simulation. The system was divided into two layers, following the ONIOM subtractive scheme. The high-layer consists of the glucosamine residue that will be sulfated in the reaction catalyzed by 3-OST, the phosphosulfate termination of PAPS and the side chains of six residues involved in the reaction or in strong hydrogen bonds with the reacting system (Glu184, His186, Asp189, Lys162, Lys215 and Lys368). The positive lysine residues balance the active site negative charge [36]. The high-layer has 83 atoms (shown in Figure 11) and was treated with density functional theory (DFT) at the B3LYP/6-31G(d) level for geometry optimizations [84, 85], whereas the rest of the system was treated at the MM level, using the AMBER force field.

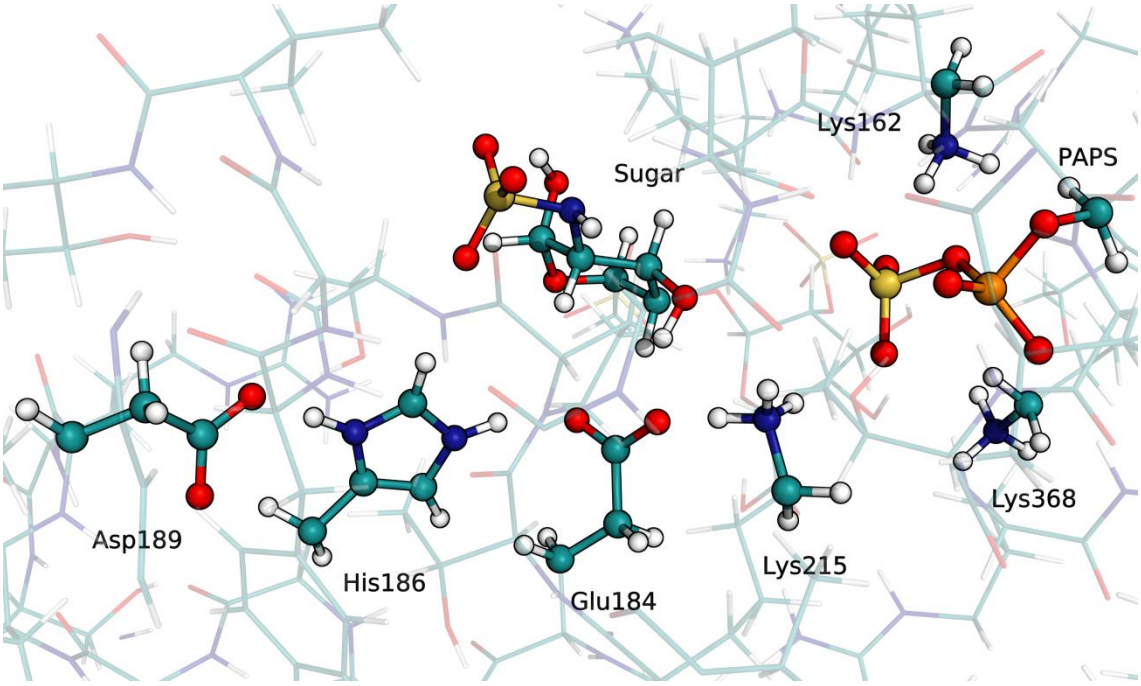

Figure 11 - Representation of the high-level layer from the model used in the QM/MM studies for 3-OST. From left to right, fragments of Asp189, His186, Glu184, Lys215, Lys162 and Lys368 are represented. The glucosamine fragment of the disaccharide present in the model and the phosphosulfate end of the co-substrate PAPS can also be seen in this figure. This layer was treated with DFT at the B3LYP/6-31G(d) level.

All the calculations were run using the Gaussian 09 software [77]. The partial charges from the MM region were included in the QM Hamiltonian (Electronic Embedding), allowing for an accurate description of the electrostatic interaction between both layers and polarization of the QM layer by the MM layer. Since the reaction catalyzed by 3-OST is a chemical reaction which does not involve large changes in structure or solvation, a Potential Energy Surface (PES) profile is a viable approach in order to accurately study the catalytic mechanism, by correlation with previous experimental studies which have already been mentioned. This method requires the definition of a geometric coordinate, such as a bond length or a dihedral angle, in order to model the reaction. This coordinate is then incremented by a fixed value and the structure is optimized in the different values of the coordinate, from reactants to products. Several reaction coordinates were tested, with the best results being obtained when the distance between the sulfate S atom and the glucosamine 3-O atom was diminished successively. The free system was fully optimized for every different reaction coordinate value, therefore building a potential energy profile from the reactants to the products. The energy maximum obtained from the PES along the reaction path corresponded to the approximate Transition State (TS) geometry. The increment was -0.05 Å in each step, decreasing to -0.01 Å once the TS geometry was nearly achieved. Subsequently, the TS structure was freely optimized, and the vibrational frequencies were calculated, to confirm the existence of a single imaginary frequency, to determine

the zero point energy (ZPE), and to calculate the entropic and thermal contributions to the free energy. Afterwards, we manually moved the reaction coordinate slightly towards the reactants and products and fully optimized these two structures. The stationary points were used to determine the activation and reaction free energies of the catalytic mechanism.

Single-point energy calculations were performed on the optimized geometries of reactants, TS and products, using different density functionals and a more complete basis set (6-311++G(2d,2p)). The density functionals tested were B1B95 [86] M06 [87] and M062X [87], BMK [88], CAM-B3LYP [89], B98 [90, 91], wB97XD [92], PBE1PBE [93, 94] and mPW1PBE [93, 95]. Dispersion corrections were applied to the tested functionals [96].

This protocol applies both to the investigation of the enzyme catalytic mechanism, as well as the exploration of the enzyme conformational space. For this study, several structures from the MD simulations were extracted, and the QM/MM protocol was repeated for each one. This will allow the investigation of the influence of the enzyme conformation in the activation barrier for the reaction.

1.4. Virtual Screening

In order to identify potential compounds with inhibitory activity towards 3-OST, a small, in-house assembled library containing 226 natural compounds and a collection of 20 compounds which show β-arylsulfotransferase-IV (β-AST-IV) inhibitory activity [97, 98] were used to perform a virtual screening protocol. The natural compounds belong to several different structure classes. This type of natural products have traditionally played an important role in drug discovery, being at the base of most early medicines [99-101]. The enzyme model was the one used in the model C of the MD simulations. The reason for this has to do with the results of the pK_a estimations and will, therefore, be expanded in later chapters. Several grid maps were tested, with the one which performed better when the protocol was validated being a 59x46x97 point grid (X,Y,Z), with 0.375 Å spacing between points, with the grid centered approximately at (30, 76, 19) coordinates. This grid comprises the key catalytic residues and substrate molecules, with additional space to accommodate for the flexibility of the docked substrate. The Lamarckian genetic algorithm [102] was employed with the following parameters:

- Population size: 150

- Maximum number of energy evaluations: $2.5 * 10^6$

- Maximum number of generations: 27000

- Number of solutions per compound: 50

Validation of the protocol was made by comparing the native substrate (in this case, the tetrasaccharide) with the docked substrate; if the RMSd value was acceptably low and the substrate was docked in a position such as to allow the catalytic reaction to occur, the protocol was considered to be validated.

2. 2-*O*-sulfotransferase

2-*O*-sulfotransferase catalyzes the step immediately before the one 3-OST does, in the HS biosynthesis. As such, this enzyme may also be a potential target for the development of drugs which prevent HSV-1 entry into the cells. Furthermore, and as has already been mentioned for 3-OST, this enzyme is also a member of the sulfotransferase, which means it, too, can provide important insight into the catalytic mechanism, as well as the structure and behavior, of the sulfotransferase family in general. Similarly to what has been described for 3-OST, the first step of this study involves building an accurate model which mimics the biological conditions and, as such, it is necessary to obtain the crystallographic structure of the enzyme from the PDB. The key residues for 2-OST are positively charged, which, although presenting differences from 3-OST, makes it so that the protocol for this enzyme is very similar to the previous one. Mutagenesis and structural studies provide the knowledge of the key active site residues, whose protonation state has to be evaluated through web-based prediction tools, as well as confirmed by MD simulations. However, although very similar to 3-OST, this enzyme presents some differences which should be taken into account. Since the key residues are positively charged, there is no need to add more residues to the high level QM layer in order to stabilize the substrate negative charge. The major differences will, therefore, lie in the model building, since the residues will be different. During this chapter, the steps which have already been detailed for 3-OST will be mentioned but not expanded upon; in those cases, the protocol was identical for both enzymes.

This work is still in progress, and therefore not all the steps that were performed on 3-OST were able to be done for 2-OST, due to time constraints.

2.1. Building the model

The model used for this study was obtained from the 2-OST X-ray crystallographic structure with the Protein Data Bank (PDB) [76] ID 4NDZ (resolution: 3.45 Å). There are two structures corresponding to 2-OST (PDB IDs 4NDZ and 3F5F), with some minor, but significant, differences. Both are fused with maltose-binding protein and both are from *gallus gallus*, contrary to 3-OST. The one that was used throughout this work is more recent (2014, compared to 2008) and from the same group as the older one. Furthermore, although it presents a lower resolution when compared with the older one (3.45 vs 2.65 Å), it contains the full trimer for the enzyme, instead of only one unit. Lastly, the more recent one contains the substrate molecule (HS heptasaccharide), which will be sulfated during the catalytic reaction. Since there is no concrete information concerning the behavior of the heptasaccharide, only the two outermost units were removed from the model, leaving a pentasaccharide which retains the essential function of the heptasaccharide. This also served to save some computational time, since this molecule needs to be parameterized using Hartree-Fock methods, which has a significant computational cost. Similarly to 3-OST, a sulfate group also had to be modeled onto PAP. Of the six chains present in the crystallographic structure, chains B, C and D were left. A representation of the trimer can be seen on Figure 12. Even though the distance between the active sites suggests that each individual unit behaves independently, leaving the biological trimer serves as assurance that the reaction closely mimics what happens in the natural environment. The amino acid sequence from the structure used in this work (UniProt ID Q76KB1) was compared to the human sequence (UniProt ID Q7LGA3), through the use of the comparison tool in UniProt [103]. Both sequences showed a 94.1% sequence similarity, which indicates that the *gallus gallus* structure, used throughout this work, is a good indicator of the enzyme behavior in humans.

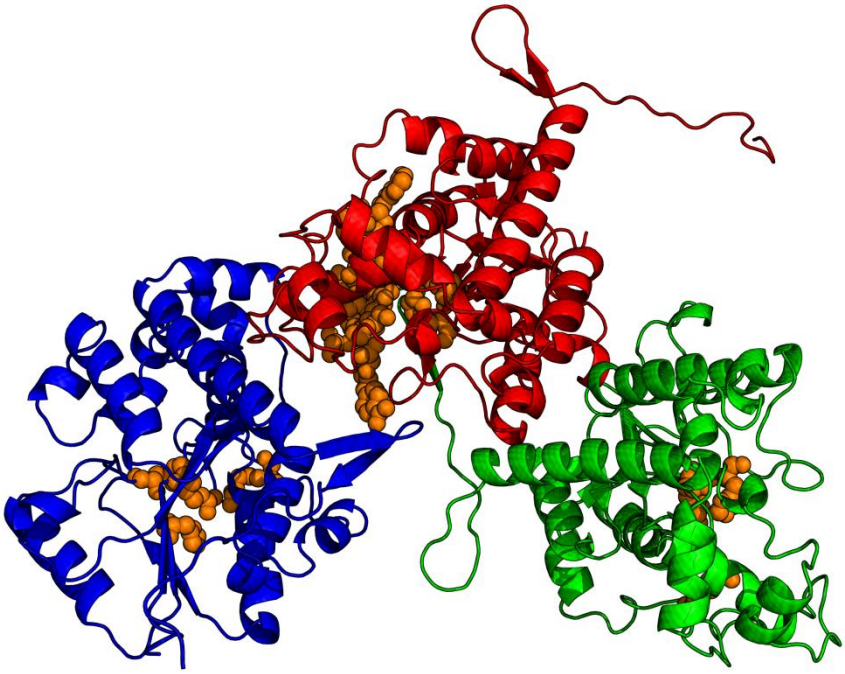

Figure 12 - Representation of the trimer structure, with each chain colored differently (green, blue and red). The chain colored red is the one that contains the substrate molecule and, as such, was used to perform the QM/MM calculations. The active site residues and substrate are colored orange.

Previous experimental and structural studies, as well as comparison with other sulfotransferases, suggest that there are four key residues for the catalytic reaction, by participating directly or indirectly in the catalytic reaction: Arg80, His140, His142 and Arg288.

2.2. pK_a estimation

Similarly to what was described for 3-OST, the estimation of the pK_a of key residues was done using a web-based prediction tool (H++), under the conditions previously described. Since the active site of this enzyme presents fewer doubts, the MD simulations were used only to equilibrate the system, due to the modeling that was made onto the enzyme. The protocol for the MD simulations was identical to the one used for 3-OST. The final atom count was 188058 and the model used had neutral His140 and His142 and positively charged Arg80 and Arg288.

2.3. QM/MM model and calculations

As was the case with 3-OST, the QM/MM calculations for 2-OST were run using the Gaussian 09 software. The initial structure was obtained from the equilibrated MD simulation. In this case, the division of the two layers was different, with the high QM layer being the phosphosulfate termination of PAPS, the glucuronic acid which will be sulfated and the side chains of the four aforementioned key residues (Arg80, His140, His142, Arg288), in a total of 77 atoms (shown in Figure 13). The rest of the protocol is similar to the one described for 3-OST, excluding the conformational analysis, which was not made due to time restraints.

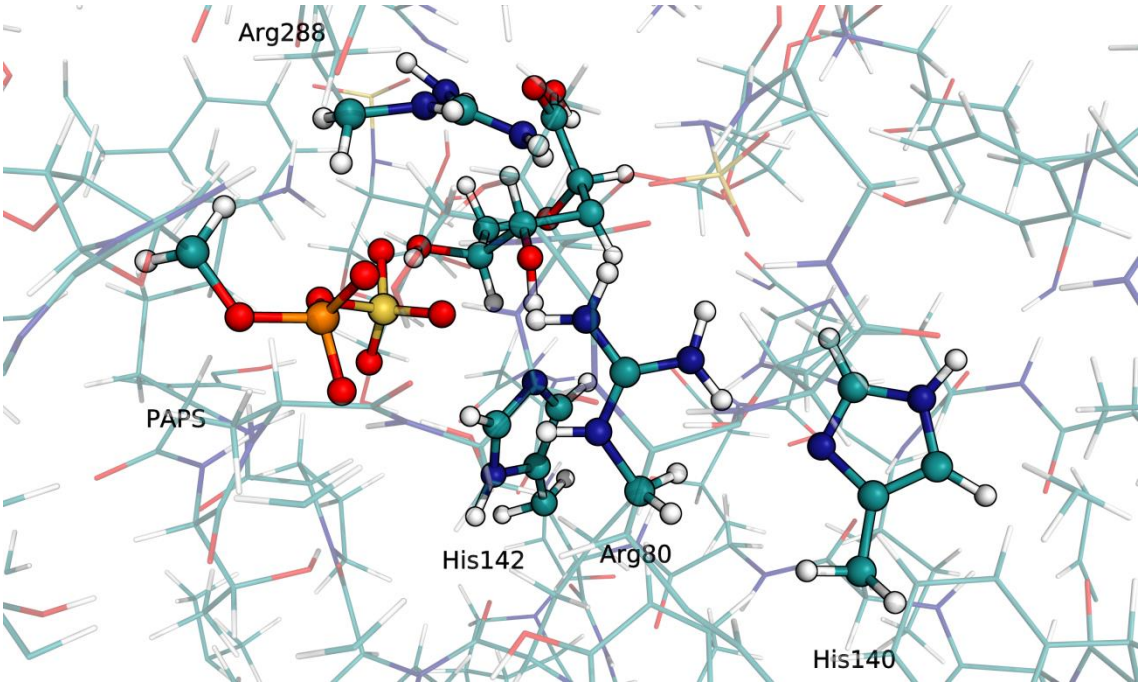

Figure 13 - Representation of the high-level layer from the model used in the QM/MM studies for 2-OST. From left to right, fragments of Arg288, His142, Arg80 and His140 are represented. The glucuronic acid fragment of the pentasaccharide present in the model and the phosphosulfate end of the co-substrate PAPS can also be seen in this figure. This layer was treated with DFT at the B3LYP/6-31G(d) level.

IV. Results and Discussion

1. 3-*O*-sulfotransferase

1.1. pK_a estimation

The crystallographic structure (PDB ID 1T8U) was used to compute the pK_a of the catalytic residues (Glu184, His186 and Asp189) in the web-based prediction tool, H++. According to this, the estimated pK_a for Glu184 is <0, for His186=6.8 and for Asp189<0. This would mean that the Glu184 and Asp189 would be deprotonated and the His186 would be mostly doubly protonated. However, since this calculation does not take into account the ligand (due to limitations of the tool itself), further confirmation is necessary.

The initial minimization protocol performed on the 3-OST crystallographic structure was done in order to minimize the effects that could arise due to the addition of the sulfate group to PAPS, which was not present in the initial X-ray structure. Three different hypotheses for the protonation state of the catalytic residues, at the reactant state, were tested by Molecular Dynamics simulations, as previously described.

Root Mean Square Deviations (RMSD) were calculated for the protein backbone in each of the models, throughout the MD simulations. As shown in Figure 14, RMSD values obtained for the three models are small (models A, B and C have average RMSD values of 2.3 Å, 2.4 Å and 2.4 Å, respectively). These RMSD values were obtained upon averaging in the time span 5-10 ns. The systems seemed equilibrated at this stage, despite a small raise in the RMSD value. The RMSD values of the active site residues and substrate molecules were also calculated and shown in Figure 15. As observed, this region also seemed equilibrated after 5 ns and, in general, the RMSD values are small. The model A has an average value slightly higher than the one obtained for the entire protein, but the opposite occurs in models B and C (models A, B and C have average RMSD values of 3.0 Å, 1.8 Å and 1.6 Å, respectively). A visual comparison between structures extracted from the three MD simulations and the crystallographic structure was made, by superimposition, as can be seen in Appendix A. In this figure, no differences in the actual folding and key residues can be discerned, whereas some differences in the sugar molecule are found.

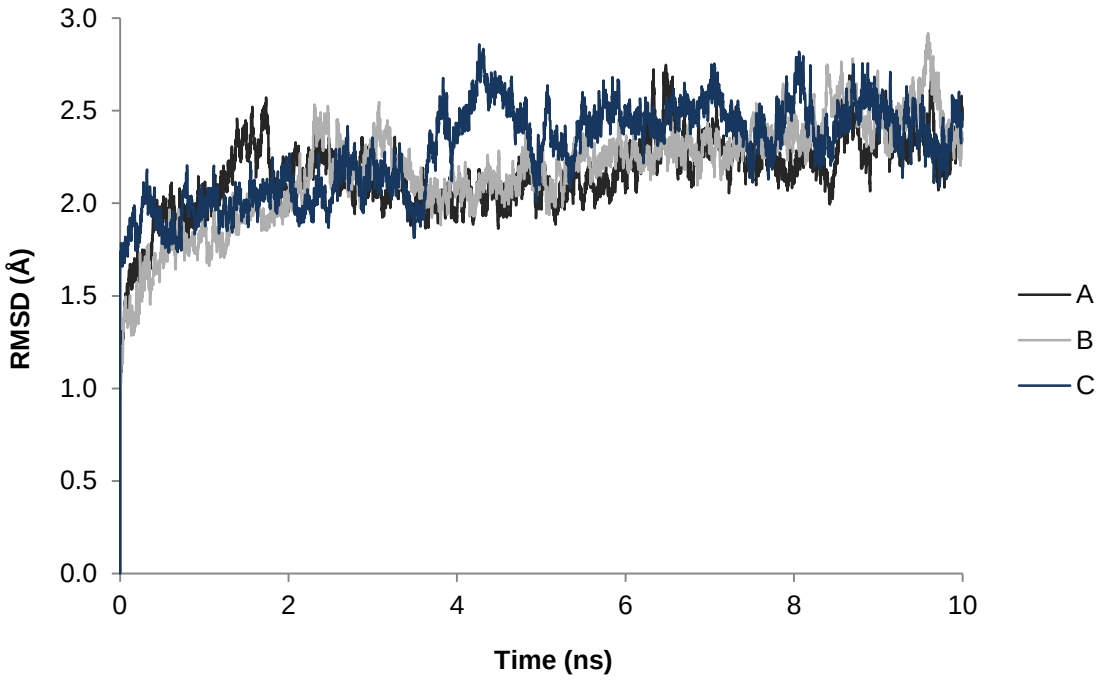

Figure 14 - RMSD of the protein backbone of the three different models throughout the MD simulations.

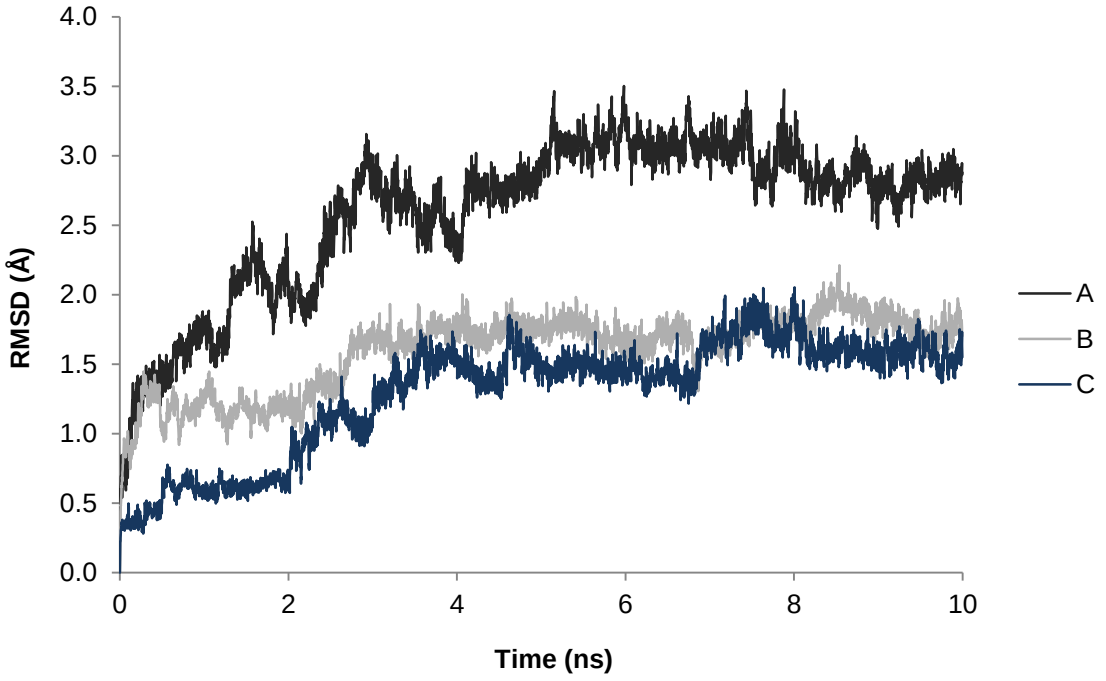

Figure 15 - RMSD of the active site residues and substrate molecules of the three different models throughout the MD simulations.

Figure 16 shows the geometry of the active site residues at the beginning and the end of each MD simulation. As shown in Figure 16, for model A (neutral Glu184 and His186, negative Asp189), the sugar molecule moved away from the PAPS molecule.

The key active site residues changed quickly to a conformation in which the catalytic reaction was not possible to occur. For model B (neutral His186, negative Glu184 and Asp189), the His186 side chain rotated, moving away from Asp189. Furthermore, the sugar and PAPS molecules moved apart throughout the simulation. Model C (positive His186, negative Glu184 and Asp189) presented the most relevant results, since all the key residues positions were kept constant along the simulation.

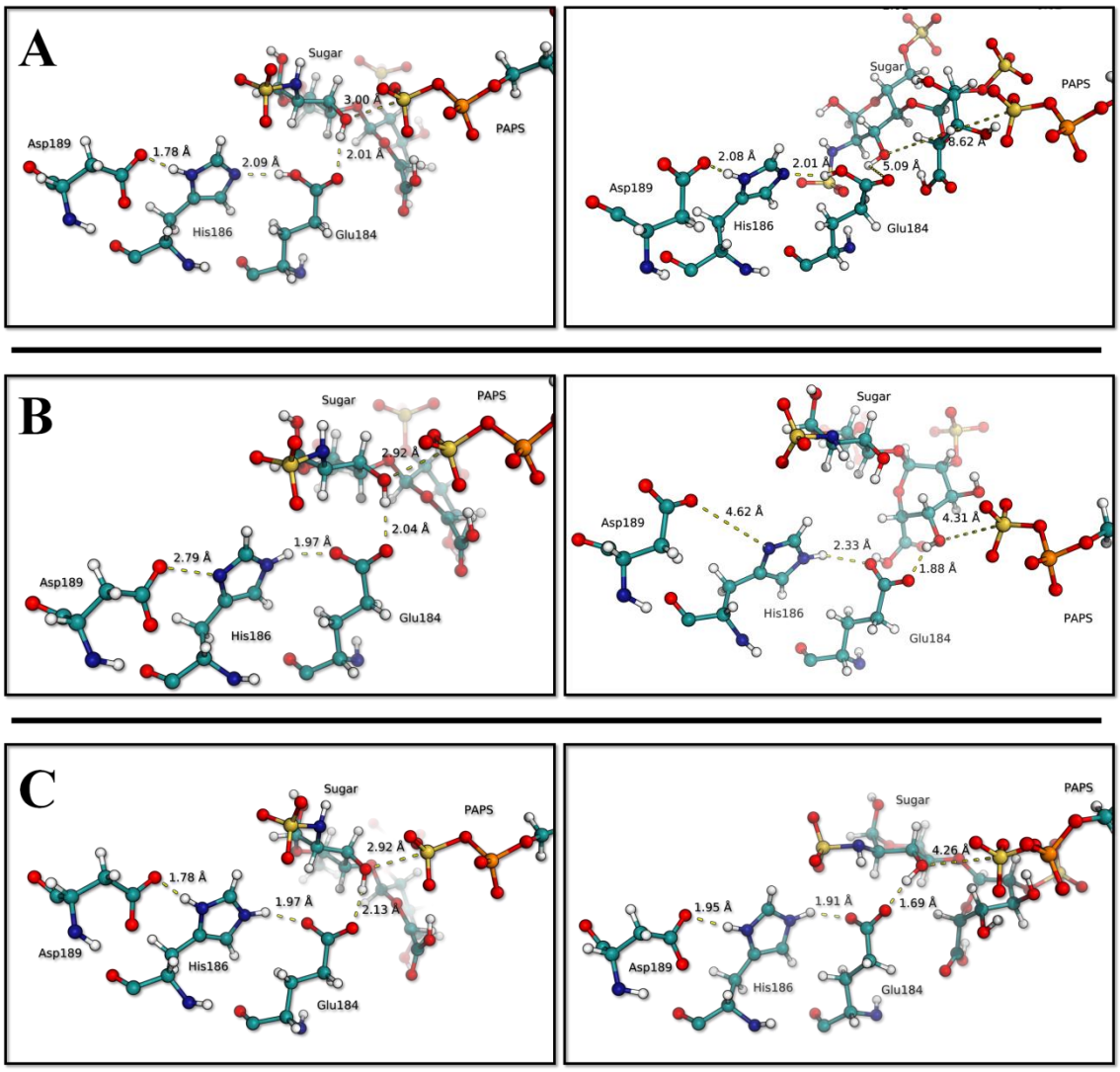

Figure 16 - Representation of the structures at the beginning (left) and end (right) of all tested MD simulations.

Table 1 represents the average distances and standard deviation (in Å), between the key residues/molecules in the active site, throughout the MD simulations.

Table 1 - Average distances and standard deviation between the key residues/molecules in the active site, throughout the MD simulations, when taking into account the 5-10 ns interval.

A	**Glu184-His186**	**His186-Asp189**	**PAPS-Sugar**	**Sugar-Glu184**
	3.82 ± 0.98	1.96 ± 0.29	8.58 ± 0.23	5.36 ± 0.64
B	**Glu184-His186**	**His186-Asp189**	**PAPS-Sugar**	**Sugar-Glu184**
	4.29 ± 1.77	7.08 ± 1.55	5.42 ± 0.31	4.56 ± 0.45
C	**Glu184-His186**	**His186-Asp189**	**PAPS-Sugar**	**Sugar-Glu184**
	2.06 ± 0.69	1.99 ± 0.34	4.84 ± 0.29	1.91 ± 0.27

It is somewhat evident, when looking at table 1, that both model A and model B seem to fail, in regards to the key distances between the key components of the active site. For model A, the average distance between PAPS and the sugar and between the sugar and Glu184 is too high, which hinders the catalytic reaction. In the case of model B, the hydrogen bond network is broken, as can be seen by the distances between Glu184 and His186, and between His186 and Asp189. For model C, all the distances remain what was expected, which further strengthens the conclusion that this model is the correct one for the study of this catalytic mechanism.

Therefore, the only protonation state of the active site residues that is viable for catalysis has a negatively charged Glu184, a positively charged His186, and a negatively charged Asp189. The result obtained in the MD simulations is in agreement with the previously mentioned pK_a estimation results. These protonation states discard the previously mentioned charge relay system hypothesis, since Glu184 has to be negatively charged and therefore cannot transfer a hydrogen atom to His186. This protonation state makes sense from a chemical point of view as it clearly maximizes the electrostatic interactions between the three residues. The two negative charges around His186 will raise its pK_a, promoting its protonation.

As a means to get further assurance that these protonation states were correct, the proton was moved back from His186 to Asp189 and the system was reoptimized at the ONIOM(B3LYP/6-31G(d):AMBER) level. The proton moved back to His186 immediately and spontaneously. The same happened when the proton was moved from His186 to Glu184. Thus, it can be concluded that the enzyme favors the "double salt bridge" configuration, with a positive His salt-bridged to two negative carboxylates, which makes sense from an electrostatic point of view.

1.2. QM/MM model and calculations

The QM/MM model described in the previous chapter, extracted from the MD simulations (at 8 ns), was optimized for the convergence criteria on the Gaussian 09 software, with the high layer being treated with B3LYP/6-31G(d) and the low layer with the AMBER force field. The boundary between both layers was dealt with by the addition of hydrogen link atoms. The electrostatic interaction between them was treated with the electrostatic embedding method, which includes the MM point charges in the QM hamiltonian. The resulting structure, which can be seen on Figure 17, only shows minor differences, namely in the positions of Lys368, whose side chain is not facing the PAPS group and in the protonation state of Lys215 and Glu184. The core structure of the active site remains intact, with the hydrogen bond network between Asp189, His186, Glu184 and Lys215 remaining established. Furthermore, the RMSD value for the protein backbone, between the X-ray structure and the optimized QM/MM geometry is small (1.3 Å), which indicates that the overall folding of the enzyme was preserved and that the model that was built is adequate for the study of the catalytic mechanism of 3-OST.

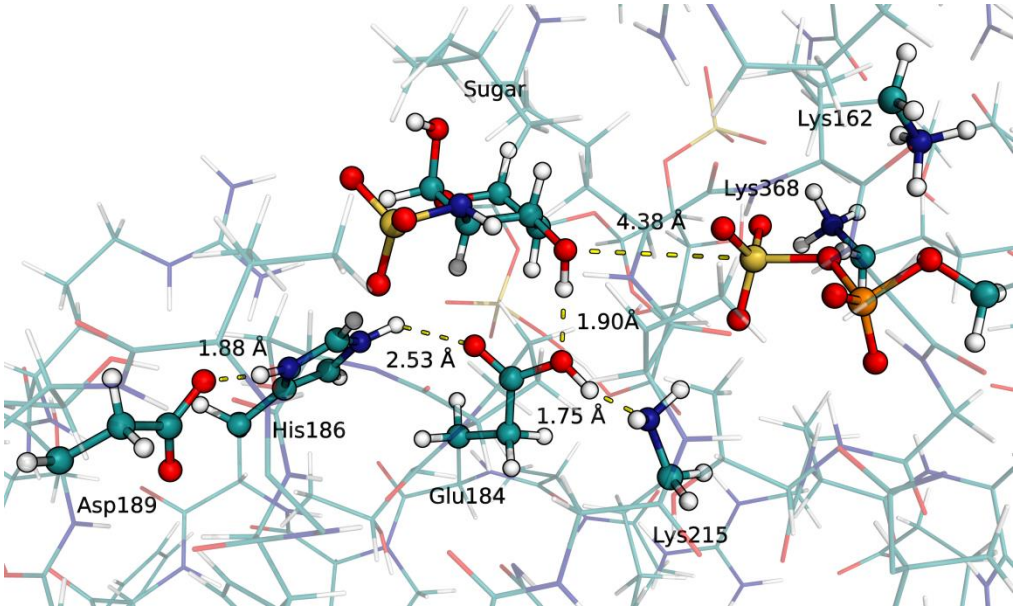

Figure 17 - Resulting structure after optimization at the B3LYP(6-31G(d)):AMBER level and interactions established at the active site.

As can be seen on Figure 17, Asp189 makes a strong hydrogen bond with His186 (1.88 Å), which in turn makes a weaker hydrogen bond with Glu184 (2.53 Å). Glu184 is now protonated, having received a proton from Lys215. Both residues establish a hydrogen bond (1.75 Å). PAPS and sugar are 4.38 Å apart, whereas the sugar is

positioned 1.90 Å from Glu184. An important thing to note is the protonation of Glu184, as well as the deprotonation of Lys215, which had not been observed previously. With the optimized structure it was, then, possible to start the study of the catalytic mechanism of 3-OST. In order to do that, as was previously described, a PES scan was performed, with the shortening of the S(PAPS)-O3(Sugar) bond as the reaction coordinate, 0.05 Å each scan step (0.01 Å near the TS). This catalytic mechanism only involves the transfer of a sulfate group and a proton (from the sugar to Glu184), since the charge relay mechanism proposed in the literature has been disproved through the study of the residues pK_a; this means that it is likely that the reaction happens through a single step, similarly to other sulfotransferases [104]. After the full reaction path was obtained, the stationary points corresponding to the reactants, transition state and products structure were freely optimized, with force constants calculated on the first point of optimization (reactants and products) or at every point of optimization (transition state). The next section will highlight the results obtained from this step.

1.2.1. Reactants

In the structure of the reactants, the sulfur atom is positioned at 3.92 Å from the 3-O position on the glucosamine unit (O3). The O3 proton is hydrogen-bonded to the Glu184 carboxylate (1.89 Å), which in turn is hydrogen-bonded, albeit weakly, to His186 (2.59 Å) and to Lys215 (1.82 Å). The former makes a short H-bond with Asp189 (1.82 Å), as can be seen in Figure 18.

The active site negative charge is stabilized by the presence of three lysine residues (Lys162, Lys215 and Lys368). These residues do not directly participate in the catalytic reaction, but they form hydrogen bonds with both PAPS and the catalytic Glu184. The protonation state of the catalytic residues has changed during the optimization, relatively to what was previously mentioned: His186 remains protonated, Asp189 remains negatively charged and Glu184 remains neutral as mentioned for the optimized structure, due to the transfer of a proton from Lys215. However, this change in protonation for the reactants structure should have no influence in the overall reaction. Figure 18 shows the main optimized catalytic core and the most important interactions established at the reactants.

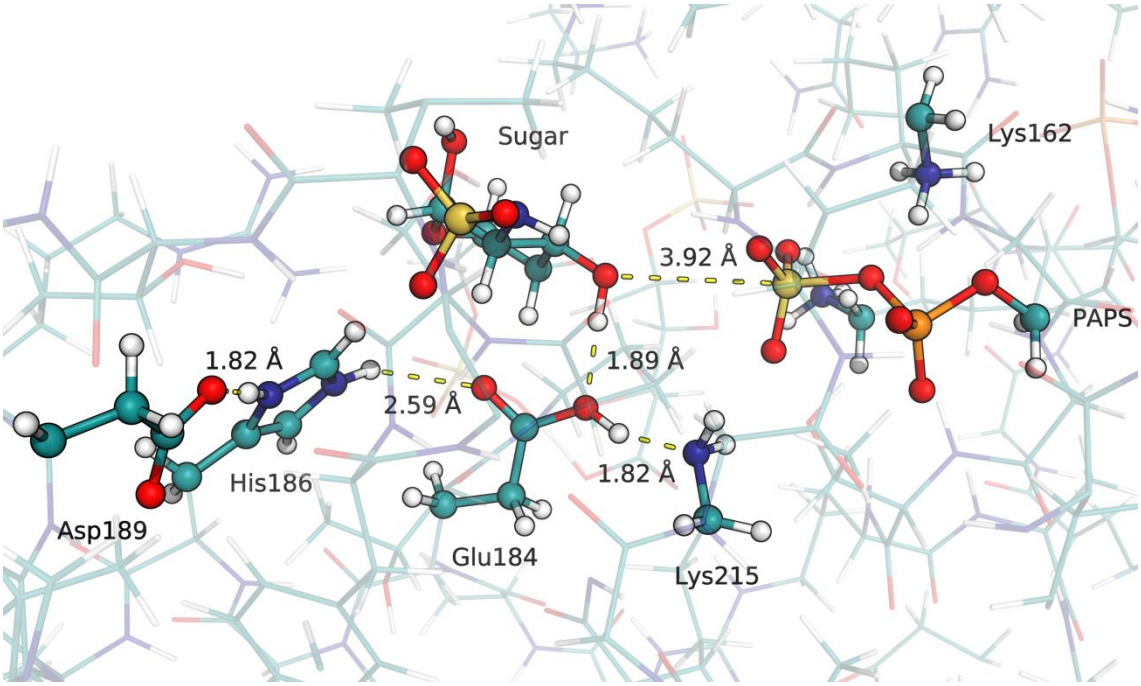

Figure 18 - Representation of the catalytic core and the most important interactions established at the reactants. Distance values are in angstroms.

1.2.2. Transition State

Figure 19 shows the main catalytic core and the most important interactions established at the transition state (TS) structure. The sulfate group is involved in an associative bipyramidal trigonal TS with a dissociative character, with the oxygen atom from the sugar and from PAPS in the axial positions. This group is positioned completely apart from PAPS, at 2.85 Å and closer to the O3 atom, at 2.41 Å, thus evidencing the aforementioned dissociative character. The leaving sulfo group leads to a higher negative charge in the phosphate group, which acts as the driving force to form the TS structure. The Glu184 and His186 have moved slightly apart, when compared to the reactants structure, from 2.59 Å in the reactants structure to 2.68 Å in the TS structure, whereas the distances between His186 and Asp189, and Glu184 and Lys215 residues remain similar. The protonation state of the key catalytic residues remains unchanged, with the O3 atom still protonated. The bond length of the proton in the O3 oxygen is only elongated from 0.98 Å to 0.99 Å at the TS. The hydrogen bond between O3 and Glu184 has shortened slightly, from 1.89 Å to 1.77 Å.

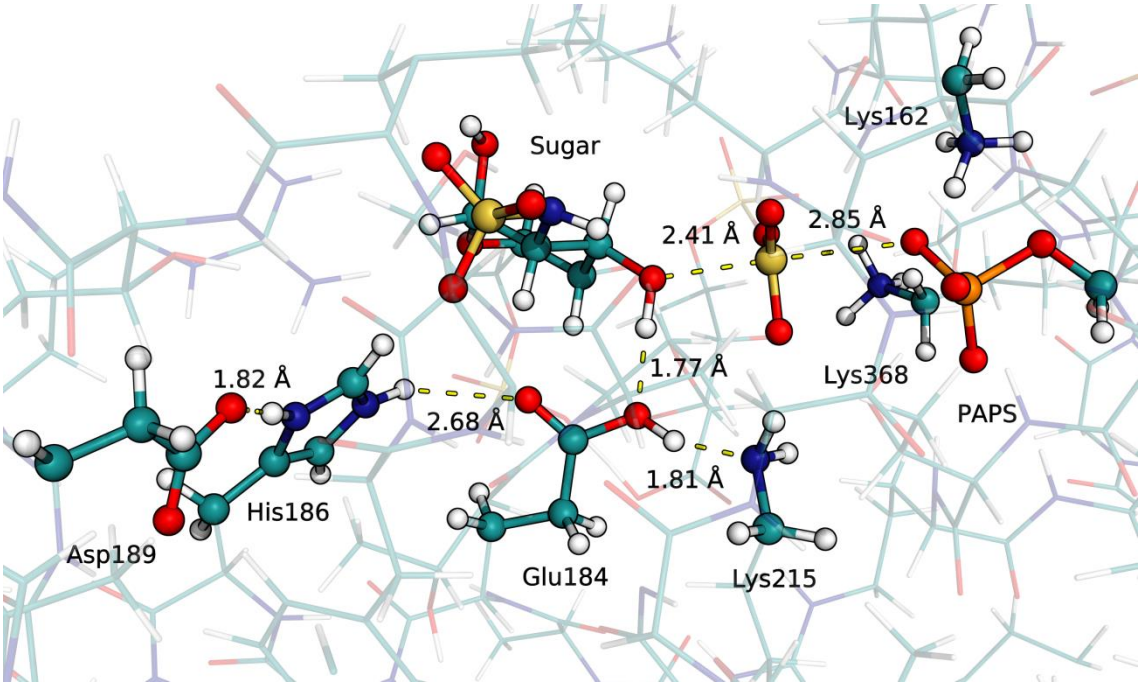

Figure 19 - Transition state (TS) structure, emphasizing the main interactions established. Distance values are in angstroms.

1.2.3. Products

The optimized products geometry is shown in Figure 20. The distance between PAPS and the sugar has increased, and the sulfur atom from the transferred sulfate group is positioned 5.21 Å away from the oxygen atom from PAPS to which it was previously bound to, and is now bound to O3 (1.70 Å), thus finishing the catalytic reaction. The proton from O3 has been fully transferred to Glu184 (0.99 Å) and is now at a distance of 1.80 Å from the O3 atom. The relative His186 and Asp189 positions and the protonation state of the active site residues remain unchanged, when compared to the TS structure. During the reaction, the proton from Glu184 was transferred back to Lys215, to accommodate for the transfer of the proton from O3. One surprise was that Lys162 also participated in the reaction. In fact, during the formation of the products, one proton from Lys162 amine group was fully transferred to an oxygen atom of the PAP phosphate group, thus stabilizing the leaving phosphate.

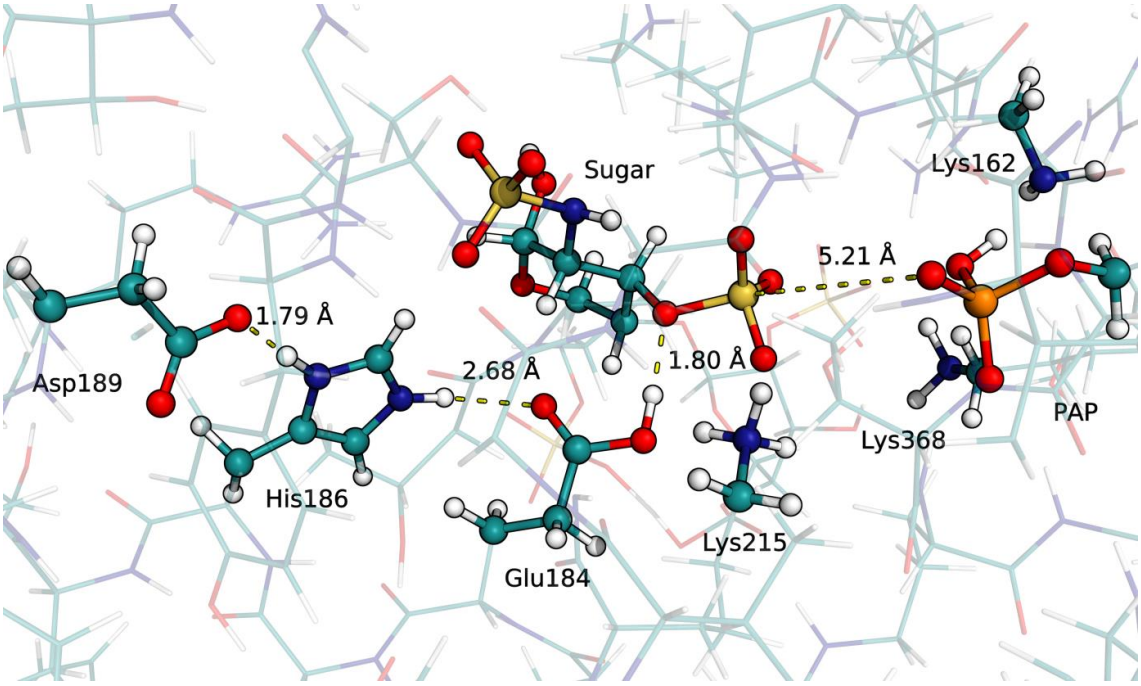

Figure 20 - Representation of the products structure, emphasizing the main interactions established. Distance values are in angstroms.

At the end of the reaction, two residues have a protonation state different from that of the reactants. Deprotonation of Glu184 and reprotonation of Lys162 has to take place before the next catalytic cycle begins. This probably occurs after product dissociation and solvation of the active site.

The study of this reaction mechanism has revealed some notable points which had not been explored up until now, namely the importance of the lysine residues in the catalytic reaction, even to the point of participating in the reaction. This is evidenced by the fact that Lys215 is deprotonated after optimization of the reactants, having to be reprotonated between one catalytic cycle and the next. Furthermore, Lys162 was also shown to have some relevance in the mechanism, albeit in a more expected way, since it stabilized the charge, through transfer of a proton, of the PAP molecule.

1.2.4. Energies associated with the mechanistic pathway

The optimization resulted in a 16.3 kcal/mol activation free energy and a -17.2 kcal/mol reaction free energy at the ONIOM(B3LYP/6-31G(d):Amber) level. Furthermore, the

reaction coordinate resulted in a smooth path connecting the reactant and product valley through the TS, which indicates that the chosen reaction coordinate is appropriate to describe this catalytic mechanism.

It is well known that the results of density functional theory have an undesirable dependence on the specific functional chosen for the calculations. To check the impact that the specific density functional has on the results we have recalculated the energy profile of the reaction using several density functionals of different types. The D3 dispersion correction was also included, and a more complete basis-set (6-311++G(2d,2p)) was used. All results from the single-point calculations are shown in Table 2. Activation free energies obtained for this reaction range from 11.2 kcal/mol (for the dispersion-corrected B3LYP functional) to 17.8 kcal/mol (for the BMK functional). Dispersion correction also appears to play a clear role in the observed barriers, lowering activation free energy when compared to their non-dispersion corrected counterparts. This decrease is less noticeable for functionals that already take into account dispersion implicitly, at least partially (such as M06, which shows a decrease of 0.6 kcal/mol, or M062X, which shows a decrease of 0.2 kcal/mol). Reaction free energies range from -24.8 kcal/mol (for the dispersion-corrected BMK functional) to -20.4 kcal/mol (for the B3LYP functional). As is the case for the activation free energy, dispersion correction also lowers the reaction free energy, although to a lesser extent.

Table 2 - Activation and reaction energies obtained for the catalytic mechanism of the 3-OST enzyme with the QM region treated with different density functionals (with dispersion corrections where applicable), using the 6-311+G(2d,2p) basis set and the electrostatic embedding scheme.

Functional	Functional Type	$\Delta G^{\ddagger}$	$\Delta G^{\ddagger}$ (D3)	ΔG_R	ΔG_R (D3)
B3LYP	h-GGA	13.7	11.2 (-2.5)	-20.4	-20.5 (-0.1)
BMK	hm-GGA	17.8	15.5 (-2.3)	-24.3	-24.8 (-0.5)
B1B95	hm-GGA	16.4	14.3 (-2.1)	-22.7	-23.1 (-0.4)
M06	hm-GGA	13.8	13.2 (-0.6)	-23.1	-23.0 (-0.1)
M062X	hm-GGA	16.3	16.1 (-0.2)	-23.8	-23.8 (=)
CAM-B3LYP	h-GGA	15.9	14.1 (-1.8)	-22.2	-22.4 (-0.2)
B98	h-GGA	14.8	-	-20.3	-
wB97XD	h-GGA	14.7	-	-22.3	-
PBE1PBE	h-GGA	16.4	14.7 (-1.7)	-21.1	-21.1 (=)
mPW1PBE	hm-GGA	16.7	-	-21.5	-

It is difficult to say, without further benchmarking tests, which should be the functional to adopt. For example, if we consider the Minnesota functionals M06 and M06-2X, which are well known to be accurate for thermodynamics and kinetics, we get a barrier of 13.2 and 16.1 kcal/mol, respectively. Previous studies describe an activation barrier,

obtained through experimental methods, of 20.4 kcal/mol (k_{cat} = 0.77 min^{-1}) [36]. For this comparison, the thermal fluctuations of the MM environment are neglected, since, in order for them to be considered, the Hamiltonian would have to be lowered; this would, in turn, lead to the introduction of a factor of error which would be comparable to the one introduced by the approximation used by the current method. This comparison is, therefore, not entirely correct, since the experimental energy is obtained by an average of several different conformations, whereas the calculated one only takes into account one conformation [105, 106]. This can lead to the dismissal of the influence that the fluctuations in the conformation of an enzyme can have in the rate of the chemical reaction (K_{cat}). This phenomenon is often called dynamic disorder and is associated with fluctuations of the rate constant, related to fluctuations of the enzyme conformation, on the same time scale as the turnover rate constant. This concern will be addressed in the next chapter, with the exploration of the enzyme conformational space, through the use of the aforementioned MD simulations and the QM/MM protocol mentioned during this chapter.

The results obtained are lower than the previously observed experimental barrier. The small differences observed in the activation barrier may be due to the fact that the sulfate group had to be modeled into the PAP molecule, therefore decreasing the overall distance from the PAPS molecule to the sugar, and due to the fact that the tetrasaccharide present in the original crystallographic structure was cleaved and only the disaccharide portion relevant to the catalytic mechanism was kept. Furthermore, there are several factors implicit in the calculations which also might have influence in the results, such as the fact that the functionals and basis sets cannot fully accurately describe the interactions occurring in the analyzed model and the fact that there is a division between two layers, with part of the interactions measured at the MM level of theory, which might also negatively translate into the results obtained. The use of non-polarizable force fields in a reaction involving highly charged species may also be a factor that impacts on the accuracy of the calculated energies. Finally, it is possible that the limiting step is not the chemical reaction. Product release is frequently rate-limiting. In the absence of further information, the experimental kinetics only represents an "upper limit" for the chemical reaction. Hence, despite all limitations that any computational calculation possesses, the overall result can be considered as accurate and within the upper limit given by experimental kinetics.

1.2.5. Exploring the Conformational Space

A good way to validate a computational enzymatic catalysis study is through comparison of the values obtained for the activation energy with the ones obtained through experimental means. However, this comparison, as was briefly mentioned in the previous chapter, can introduce an appreciable factor of error, due to the fact that computational methods take into account one conformation, ignoring the possible energy fluctuations which may be relevant in the studied time scale. As such, this validation may be flawed, due to this comparison and the fact that computational methods introduce errors in and of themselves, allowing for the obtainment of concordance between computational and experimental methods, where they might otherwise have significant differences.

There are methods that take into account the dynamic nature of the enzymatic conformations, but they come at the cost of a significant lowering of the hamiltonian complexity in order to calculate energies for a large number of structures. One approach to solve this problem, while keeping computational costs at an acceptable level, is to perform QM/MM calculations on an ensemble of different structures extracted from MD simulations. As such, one can go beyond the stationary ONIOM QM/MM method, while still keeping a high-level hamiltonian.

In order to perform this work, several structures were extracted from the aforementioned model C of the MD simulations (since it is the one which expresses the correct protonation states of the key residues), at regular 1 ns intervals, for the duration of the 10 ns. The aforementioned QM/MM protocol was then applied to these structures, in order to replicate the catalytic mechanism obtained in the previous chapter. Table 3 shows the key distances between the residues/molecules in the active site, for the converged structures obtained from the MD simulations. The Sugar-PAPS distance contains the distance from the sugar to the sulfate group and from the sulfate group to the PAP molecule, respectively, when relevant.

Table 3 - Key distances between the active site residues/molecules for the structures obtained from the MD simulations. Sugar-PAPS distance, when relevant, represent the distance between the sugar and sulfate group and from the latter and the PAP molecule. All distances are in angstroms.

0	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.72	1.81	2.58	3.59
TS	1.72	1.80	3.09	2.24 + 1.93
Products	1.78	2.45	1.94	3.57
1	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.66	1.80	2.15	3.80
TS	1.69	1.78	2.56	2.65
Products	1.63	1.95	1.80	4.22
2	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.82	1.84	1.78	3.47
TS	1.78	2.06	1.66	2.12 + 2.52
Products	1.68	3.21	1.82	3.39
3	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.99	1.73	1.85	3.53
TS	2.02	1.77	1.86	2.23 + 2.04
Products	1.96	1.56	2.20	3.84
6	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.72	1.51	1.67	3.65
TS	1.72	1.58	1.71	2.15 + 2.03
Products	1.69	1.82	1.89	4.49
7	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.83	2.53	1.90	3.18
TS	1.82	2.65	1.76	2.23 + 2.53
Products	1.80	2.68	1.80	4.51
8	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.64	1.83	1.73	3.26
TS	1.62	2.06	1.68	2.31 + 2.19
Products	1.64	1.86	2.84	4.03
10	**Asp189-His186**	**His186-Glu184**	**Glu184-Sugar**	**Sugar-PAPS**
Reactants	1.73	1.69	1.79	3.05
TS	1.70	1.85	1.77	2.10 + 2.14
Products	1.65	1.97	3.32	5.84

Although some interactions show some variation in the respective distance, most interactions remain fairly constant when comparing reactants, TS and products.

In order to obtain activation and reaction free energies, ZPE, thermal and entropic contributions were assumed to be constant and, thus, were taken from the PES scans from the previous chapter. Table 4 shows the activation and reaction free energies obtained for the converged structures extracted from the MD simulations (8 out of the 11 structures converged).

Table 4 - Activation and reaction free energies obtained for the different structures obtained from the MD simulation of model C, optimized at the ONIOM(B3LYP/6-31G(d):Amber) level. Time in nanoseconds, energies in kcal/mol.

Time	$\Delta G^{\ddagger}$	ΔG_R
0	17.8	-66.6
1	22.4	-39.6
2	20.9	-33.9
3	13.8	-19.8
6	17.7	-50.6
7	14.9	-34.4
8	13.3	-52.1
10	31.7	-35.3

Activation free energies range from 13.3 kcal/mol to 31.7 kcal/mol ($\Delta G^{\ddagger}_{exp}$ = 20.4 kcal/mol), whereas reaction free energies range from -66.6 kcal/mol to -19.8 kcal/mol. The most relevant thermodynamic quantity for the study of the enzymatic mechanism is the activation free energy, which is shown in table 4 to depend slightly on the enzyme conformation. A much larger dependence is observed for the reaction free energy. The differences among different structures can be explained by differences in the organization of the active site salt bridges between the phosphate, sulfate, and the three lysines. As these ionic interactions are very strong, a small difference in the salt bridges network gives rise to large differences in free energy. Even though this study is valuable in order to refine the free activation and reaction energy values, a much larger number of conformations would be desirable, in order to converge these values to a larger degree of certainty.

The relation between k_{cat} and the activation energy is given by the transition state theory formula, as shown in equation 14:

$$k_{cat} = \kappa \frac{k_B T}{h} e^{-\frac{\Delta G^{0\ddagger}}{RT}} \quad \textbf{(14)}$$

where κ corresponds to the transmission coefficient, k_B corresponds to the Boltzmann constant, T is the temperature, h is the Planck constant and R is the ideal gas constant. In order to calculate the observed k_{cat}, one must take into account the k_{cat} values for all the reactant structures that the enzyme can adopt. At any moment, the enzyme is at a certain conformation, to which corresponds one activation energy for the reaction to occur. The k_{cat} for that conformation can be derived from the activation energy by applying equation 14. The enzyme has a certain probability to adopt a certain conformation; however, this probability cannot be estimated. Therefore, it is reasonable to assume that every conformation has an equal probability to occur, at which point a weighted k_{cat} can be calculated from the average of all the k_{cat} obtained from the different structures. After obtaining the average k_{cat}, this value can be reconverted to activation energy, once again through equation 14. Applying this to the current study, by calculating the k_{cat} for all eight structures in table 4, one can obtain the weighted k_{cat} which can be reconverted into an observed activation energy, in this case of 14.3 kcal/mol at the ONIOM(B3LYP/6-31G(d):AMBER) level. However, the number of structures in this sampling is low, so the next task would be to obtain more structures and refine this energy.

1.2.6. Virtual Screening

The virtual screening study was performed, as has been mentioned before, in order to assess which compounds, out of the library of natural compounds and compounds with known sulfotransferase inhibitor activity, would better be able to inhibit the 3-OST activity.

In order to validate the docking protocol and parameters chosen, several different grids were attempted and the substrate was re-docked into the protein. The resulting binding pose was then compared to the crystallographic structure in order to assess if the docking had been properly done. For the grid that was described in the methods chapter, which was used in this step, the virtual screening protocol provided a solution which had an RMSD value of 3.1 Å, when considering the disaccharide which was used in the QM/MM calculations (1.9 Å when only considering the reacting monosaccharide). Although this RMSD value is somewhat high, one has to consider that the substrate is a sugar molecule, which has substituents which are very flexible. Figure 21 shows a superposition of the disaccharide used in the QM/MM calculations

and the one re-docked into the protein, evidencing the very similar binding pose of both.

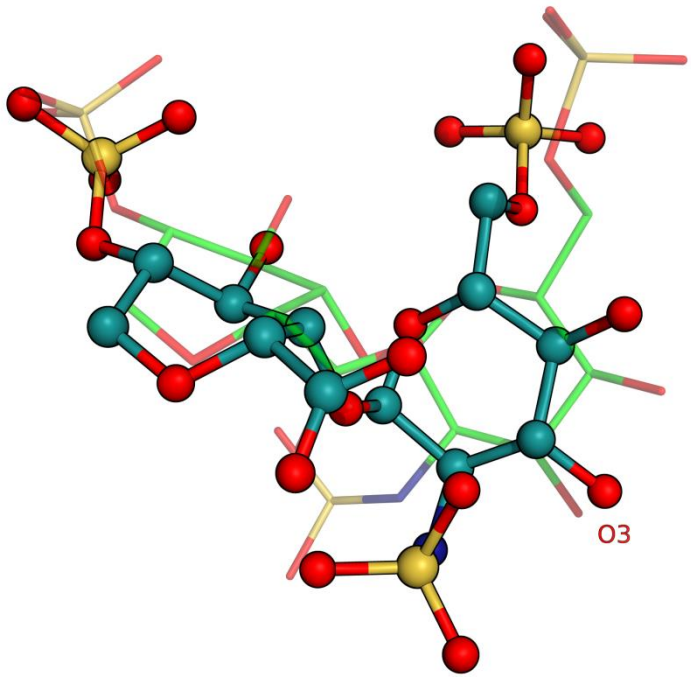

Figure 21 - Representation of the superposition of the crystallographic disaccharide with the docked disaccharide. The docked structure is represented in CPK, whereas the crystallographic one is semi-transparent and represented in lines. The O3 atom is evidenced.

The disaccharide molecule establishes important interactions with the enzyme, responsible for its binding pose. The main interactions are with Arg166, Glu184, Lys215, Lys259, Thr367 and Lys368 and can be seen in Figure 22.

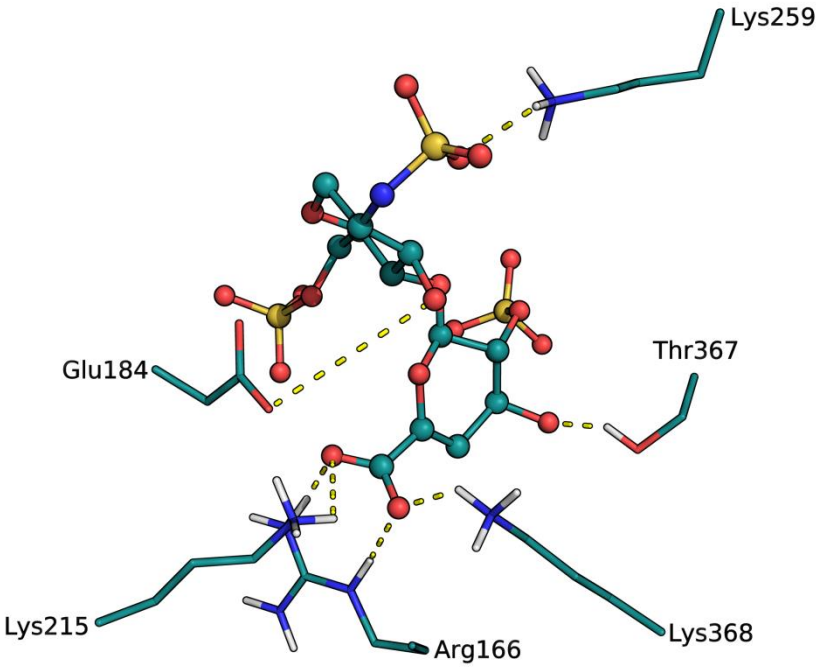

Figure 22 - Representation of the interactions between the docked disaccharide and the active site residues.

After the validation of the virtual screening protocol, several compounds (246 total) were docked into the 3-OST active site in order to rank their ability to bind to this enzyme. Table 5 shows the top five results found for the natural compounds and the top five results found for the compounds with sulfotransferase inhibitor activity, ranked based on the $\Delta G_{binding}$ to the 3-OST enzyme.

Table 5 - Top five natural compounds (A to E) and top five compounds with known sulfotransferase inhibitor activity (F to J) used in the virtual screening protocol, ordered based on the $\Delta G_{binding}$ to the 3-OST enzyme.

Name	$\Delta G_{binding}$ (kcal/mol)	Ki (nM)
A	-10.21	32.8
B	-10.12	38.1
C	-9.96	49.9
D	-9.88	57.2
E	-9.87	58.1
F	-7.25	4840
G	-7.23	5010
H	-7.04	6900
I	-6.92	8450
J	-6.71	12100

Compounds A to E correspond to natural compounds, whereas compounds F to J correspond to compounds with known sulfotransferase inhibitor activity.

As can be seen on table 5, there seems to be a tendency which indicates that the natural compounds (A-E) are better 3-OST inhibitors than the compounds with known sulfotransferase inhibitor activity (F-J). In fact, the $\Delta G_{binding}$ values show a somewhat large difference between the two groups of compounds. The figures shown next (23 to 31) show the interactions established between the enzyme residues and the top-ranked docked compounds. Residues not establishing interactions with the compound are not shown in the following figures.

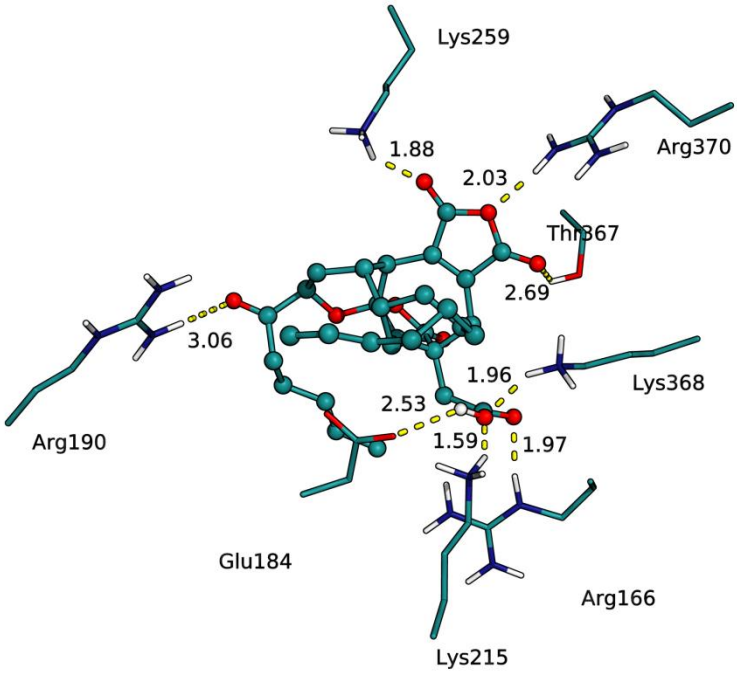

Figure 23 - Representation of the interactions established between the docked compound A and the active site residues. All distances are in angstrom.

Compound A shows extensive interactions with the enzyme, establishing strong hydrogen bonds with Lys215, Glu184, Arg166, Lys368, Arg370, Lys259 and Thr367 and a somewhat weaker strong hydrogen bond with Arg190.

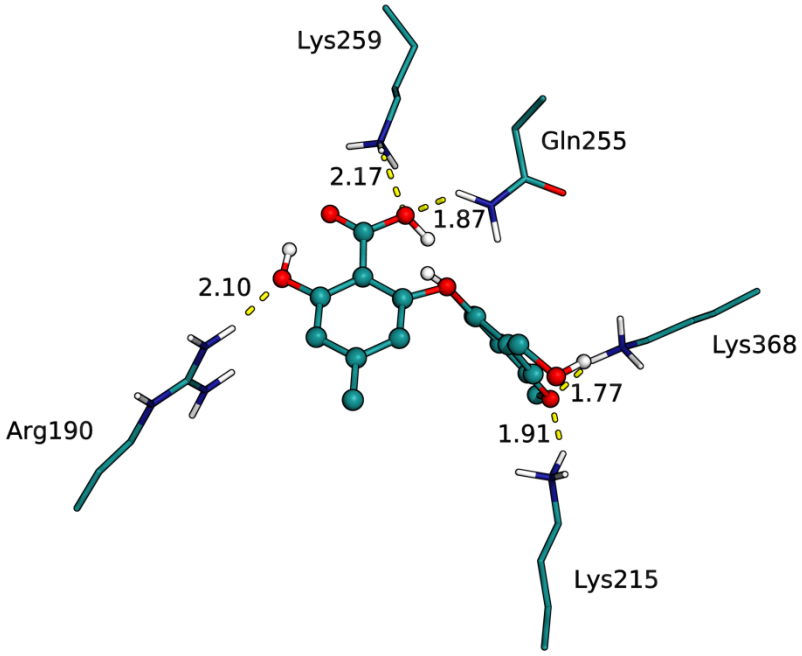

Figure 24 - Representation of the interactions established between the docked compound B and the active site residues. All distances are in angstrom.

Compound B shows interactions with fewer residues than compound A, establishing strong hydrogen bonds with Arg190, Lys259, Gln255, Lys368 and Lys215.

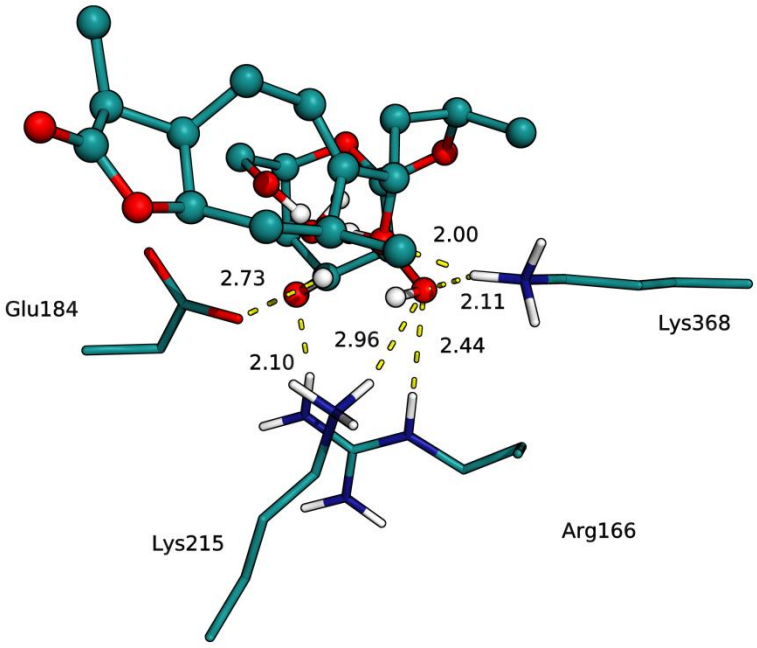

Figure 25 - Representation of the interactions established between the docked compound C and the active site residues. All distances are in angstrom.

Similarly to compound B, compound C establishes strong hydrogen bonds with Glu184, Arg166 and Lys368 and a weaker hydrogen bond with Lys215.

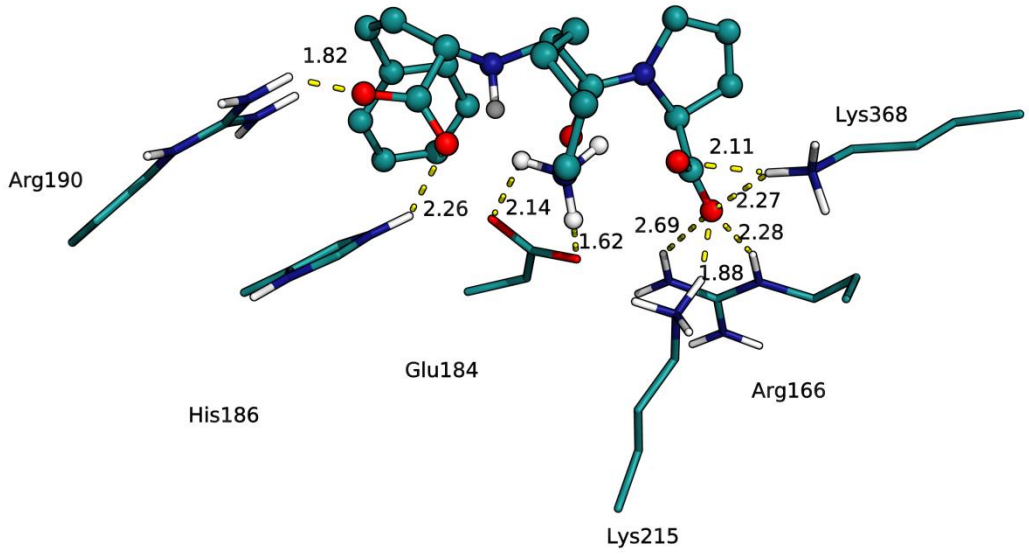

Figure 26 - Representation of the interactions established between the docked compound D and the active site residues. All distances are in angstrom.

Compound D establishes more interactions with the enzyme than compounds B and C, showing hydrogen bonds with Arg190, His186, Glu184, Lys215, Arg166 and Lys368.

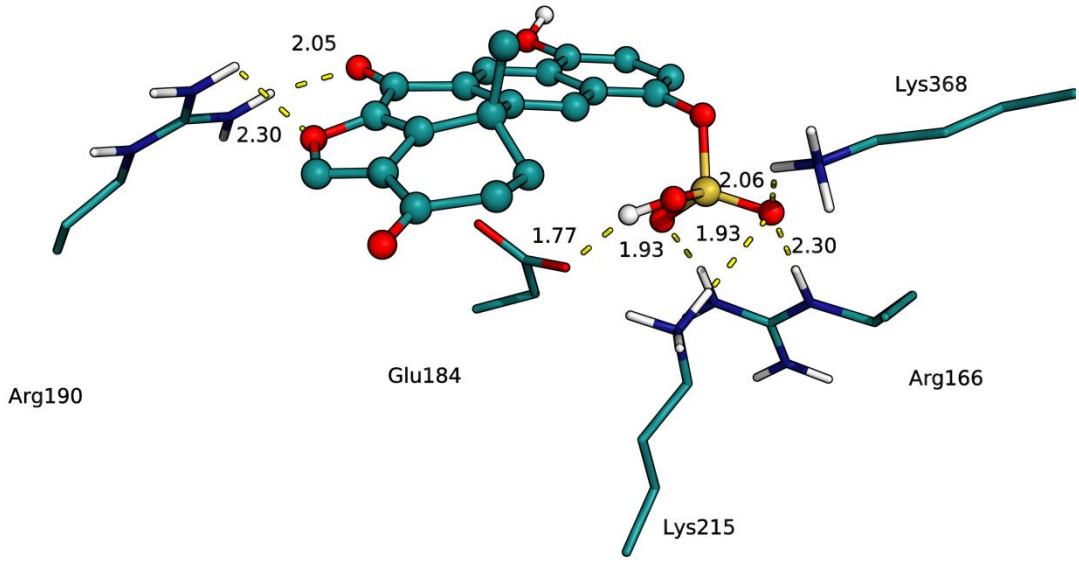

Figure 27 - Representation of the interactions established between the docked compound E and the active site residues. All distances are in angstrom.

Compound E, once again, establishes fewer interactions with the enzyme, showing strong hydrogen bonds with Arg190, Glu184, Lys215, Arg166 and Lys368. This

compound has the particularity of having a sulfate group near the region where the sugar is sulfated.

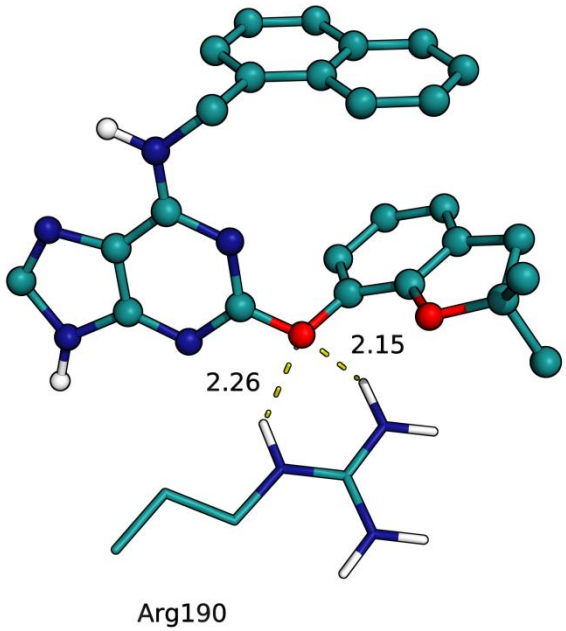

Figure 28 - Representation of the interactions established between the docked compound F and the active site residues. All distances are in angstrom.

Moving on to the compounds which show inhibitory activity towards other sulfotransferases, compound F only establishes a strong hydrogen bond with Arg190, which is unusual when comparing it with the previous compounds; this is in accordance with the "jump" in the $\Delta G_{binding}$ value that was observed in table 5.

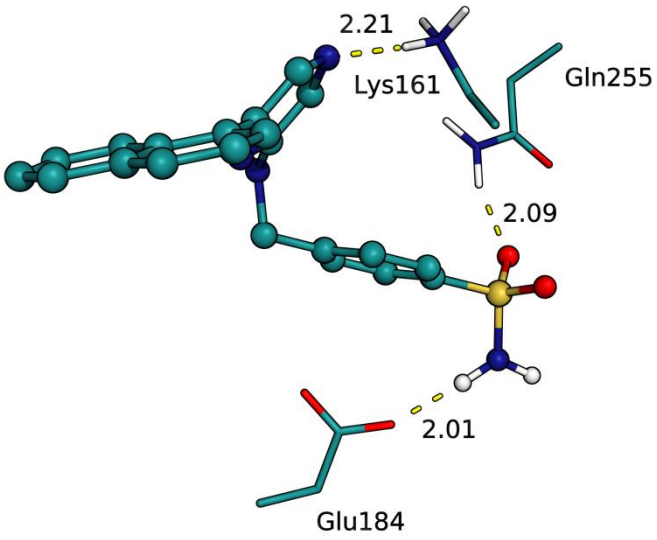

Figure 29 - Representation of the interactions established between the docked compound G and the active site residues. All distances are in angstrom.

Compound G, similarly to compound F, establishes strong hydrogen bonds with Lys161, Gln255 and Glu184.

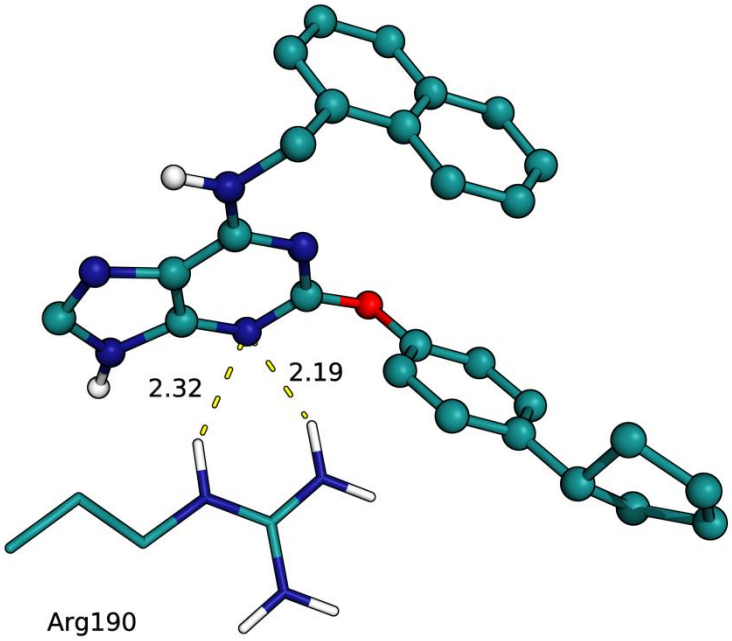

Figure 30 - Representation of the interactions established between the docked compound H and the active site residues. All distances are in angstrom.

Compound H is very similar, in structure, to compound F. As expected, it is also similar in the interactions it establishes with the enzyme, only showing a strong hydrogen bond with Arg190.

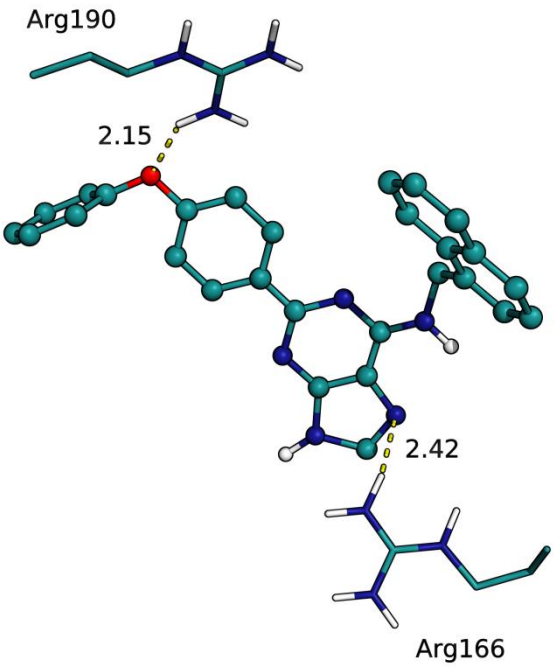

Figure 31 - Representation of the interactions established between the docked compound I and the active site residues. All distances are in angstrom.

Compound I follows the tendency of the previous compounds, only establishing two hydrogen bonds with Arg190 and Arg166.

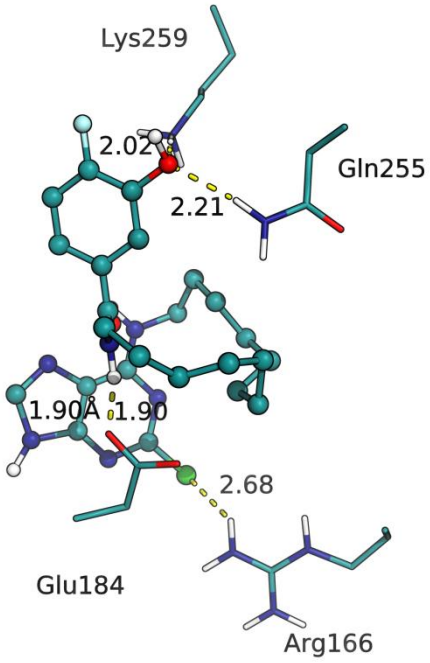

Figure 32 - Representation of the interactions established between the docked compound J and the active site residues. All distances are in angstrom.

Lastly, compound J, which is a peculiar one when compared with all the compounds shown so far, since it contains a chlorine and a fluorine atom, establishes hydrogen

bonds with four residues: Glu184, Arg166 (through its chlorine atom), Gln255 and Lys259.

Through analysis of the compounds shown here, it is possible to discern the influence of certain residues in positioning the docked compounds. Residues which have been mentioned in relation to the catalytic mechanism, such as Glu184, His186, Lys215, Lys368, due to being close to the active site, are some of the residues which establish hydrogen bonds with several compounds. However, there are other important residues which are repeatedly shown, as well, such as Arg190, Gln255 and Arg166. Lys259 also establishes hydrogen bonds with some of the bulkier compounds, such as compounds A and B.

Compound A (phomoidride B) [107], which is a fungal metabolite with squalene synthase and Ras farnesyltransferase inhibitor activity [108], seems to be the one which would better inhibit 3-OST. This is not only due to the $\Delta G_{binding}$, but also because its interactions with 3-OST are similar to the ones observed for the disaccharide. Both the compound and the disaccharide establish interactions with Glu184, Lys215, Arg166, Lys368, Lys259 and Thr367. Furthermore, compound A also interacts strongly with Arg370 and Arg190.

Compound B (Barceloneic acid A), which is also produced by a fungus and possesses farnesyl-protein transferase inhibitor activity [109], also presents a good $\Delta G_{binding}$ to 3-OST, showing interactions with it in common with the docked disaccharide. Both interact with Lys215, Lys368 and Lys259, whereas compound B also establishes interactions with Arg190 and Gln255.

Out of the compounds analyzed here, it seems that compound A and compound B (phomoidride B and Barceloneic acid A, respectively), both natural compounds, would be strong candidates for a potential 3-OST inhibitor. Further studies with these compounds should be done, in order to evaluate their kinetic properties. These relate mostly to the rate and extent of absorption after administration, their distribution inside the organism, their metabolism after they enter the system, their excretion after they perform their action and the potential toxicity to the organism they may present. These properties are commonly known as ADMET [110, 111] and are crucial in the determination of the potential of a chemical compound as a drug [112].

2. 2-O-sulfotransferase

2.1. pK_a estimation

The crystallographic structure was used to compute the pK_a of the catalytic residues (Arg80, His140, His142, Arg288) in the web-based prediction tool, H++. According to this, the estimated pK_a for Arg80 is >12, for His140 and His142 <0 and for Arg288 10.7. This would mean that the His140 and His142 would be neutral, whereas Arg80 and Arg288 would be doubly protonated. Although the concern with the limitations of the tool still remains, the protonation states of the key residues in this case do not seem to have another possibility which would be catalytically viable. Therefore, contrary to what happened with 3-OST, the MD simulations in this case were performed on one model only, corresponding to the aforementioned protonation states.

Nonetheless, the MD simulations were still necessary, in order to once again minimize the effects that could arise due to the addition of the sulfate group to PAPS, which was not present in the initial X-ray structure. RMSD was calculated for the protein backbone, throughout the MD simulations, as shown in Figure 33. In this case, the average RMSD value (taking into account the 10-20 ns timespan, when the enzyme is equilibrated) has an unusually high value (5.89 Å).

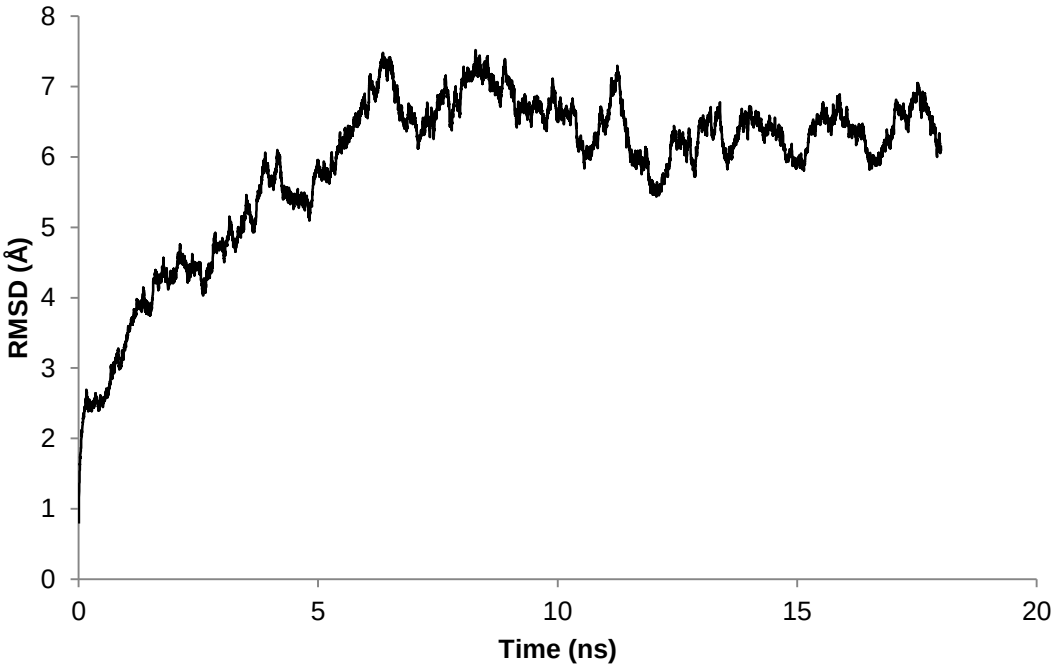

Figure 33 - RMSD of the protein backbone of the model used throughout the MD simulations.

Further investigation into this unusually high value reveals that, due to the large size of the enzyme, which is composed by three chains, and the fact that it has a significant number of flexible areas, this value is concentrated on the outer areas of the enzyme, far from the active site. Figure 34 evidences the regions of the trimer with higher RMSD values.

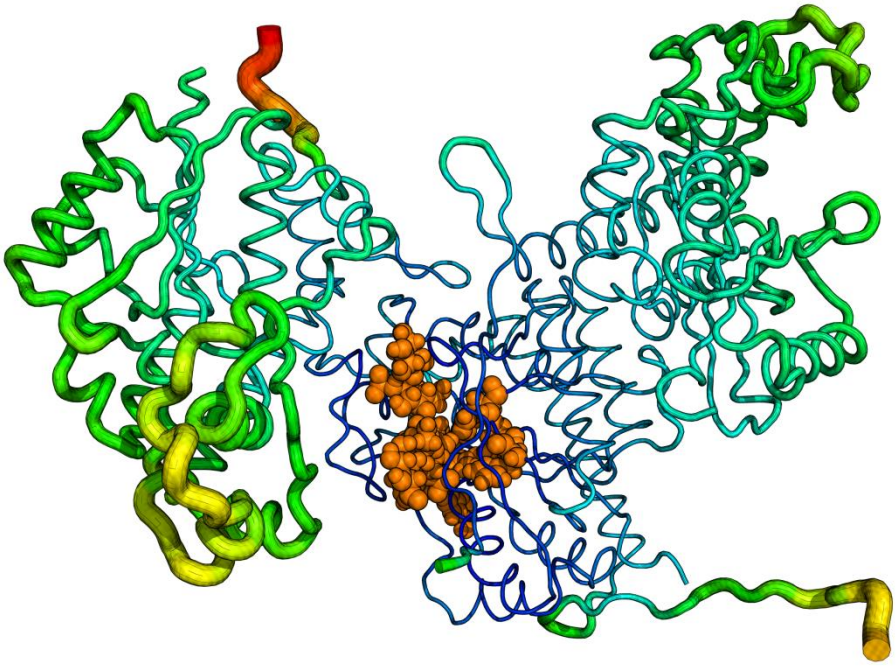

Figure 34 - Representation of the residues in the trimer based on their RMSD value. Lower values are represented by thinner cartoon cylinders and bluer colors, whereas higher values have thicker cylinders and green to red colors. The active site is represented by orange spheres (not related to RMSD value).

The RMSD values for the active site residues and substrate molecules were also calculated and can be seen in Figure 35. The average RMSD for the equilibrated structure (10-20 ns) was, in this case, 1.62 Å.

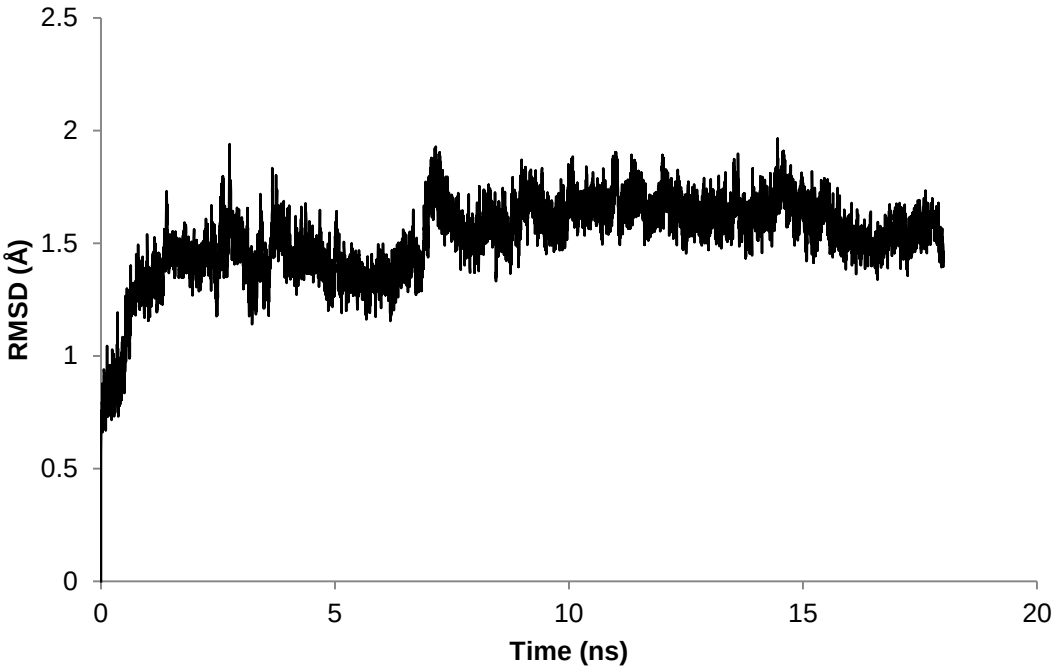

Figure 35 - RMSD values for the active site residues, throughout the MD simulations.

Figure 36 shows the geometry of the active site residues at the beginning and the end of the MD simulation. As can be seen by this comparison, the distance between the sugar molecule and PAPS remained largely similar (3.42 Å initially and 3.35 Å at the end). The distance between the OH group from the sugar and His142 was shortened from 3.84 to 1.70 Å, since the hydrogen atom rotated towards His142. The hydrogen bond formed between His140 and Arg80 became weaker, with the distance between the two residues increasing from 2.05 Å to 2.51 Å. The distance between Arg80 and the oxygen atom from the sulfo group in PAPS increased from 1.83 Å to 2.53 Å.

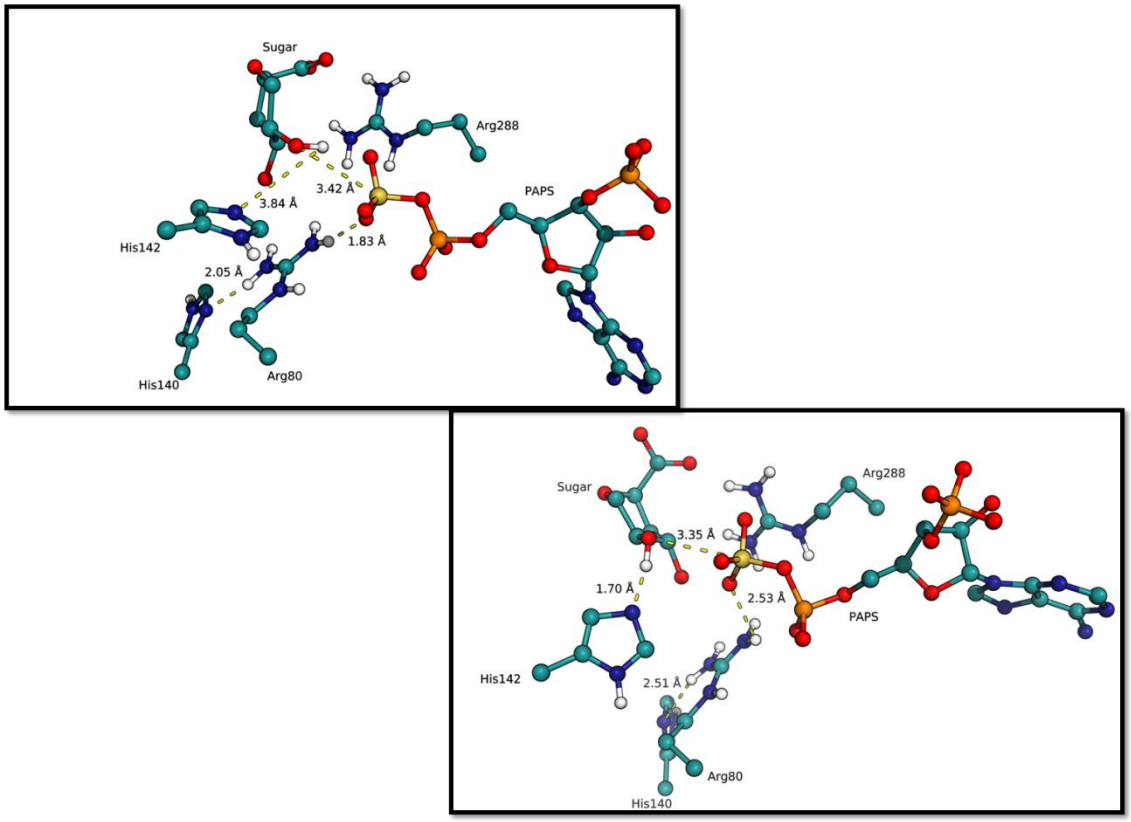

Figure 36 - Geometry of the active site key residues and substrate molecules, at the beginning (left) and the end (right) of the MD simulations. All distances are represented in angstrom.

Throughout the simulation, taking into account the point after which the structure is equilibrated, the average distance between sugar and PAPS was 3.70 Å, whereas the distance between the former and His142 averaged 3.16 Å. The distance between Arg80 and His140 averaged 2.19 Å and the one between the former and PAPS averaged 1.16 Å, indicating the presence of a strong hydrogen bond between the two.

These results indicate that the overall structure and folding of the enzyme remains intact, with the outer edges of the enzyme showing a greater deviation from the crystallographic structure, as was expected. Even so, the active site residues and substrate molecules remained similar to the crystallographic structure, with the

interactions between them staying largely unchanged. Therefore, these results strengthen the conclusion that the protonation state adopted by the active site residues is correct, in order to study the enzyme catalytic mechanism.

2.2. QM/MM model and calculations

The protocol adopted from this section was largely similar to the one described for 3-OST, with the model being obtained from the equilibrated MD simulations. This model was then optimized for the convergence criteria on the Gaussian 09 software, with the high layer being treated with B3LYP/6-31G(d) and the low layer with the AMBER force field. As previously, the boundary between the layers was dealt with by adding hydrogen link atoms, with the interaction between both layers being treated with the electrostatic embedding method. The resulting structure can be seen on Figure 37. There are, mainly, only slight differences between this optimized structure and the initial structure obtained from the MD simulations. The distance between sugar and PAPS is now 3.73 instead of 3.35 Å and the distance between Arg80 and PAPS is 2.04 Å, lower than the previously mentioned structure. The hydrogen bond between Arg80 and His140 has weakened slightly, increasing the distance from 2.51 Å to 2.65 Å. One major difference is the 2-OH position from the sugar. The proton has rotated slightly towards PAPS, distancing itself from His142 (from 1.70 Å to 3.88 Å). This means that this proton has to rotate back towards His142, in order for the catalytic reaction to occur. The overall structure and interactions at the active site level are, therefore, maintained, which is also evidenced by the low RMSD between the crystallographic structure and the optimized QM/MM geometry (1.4 Å).

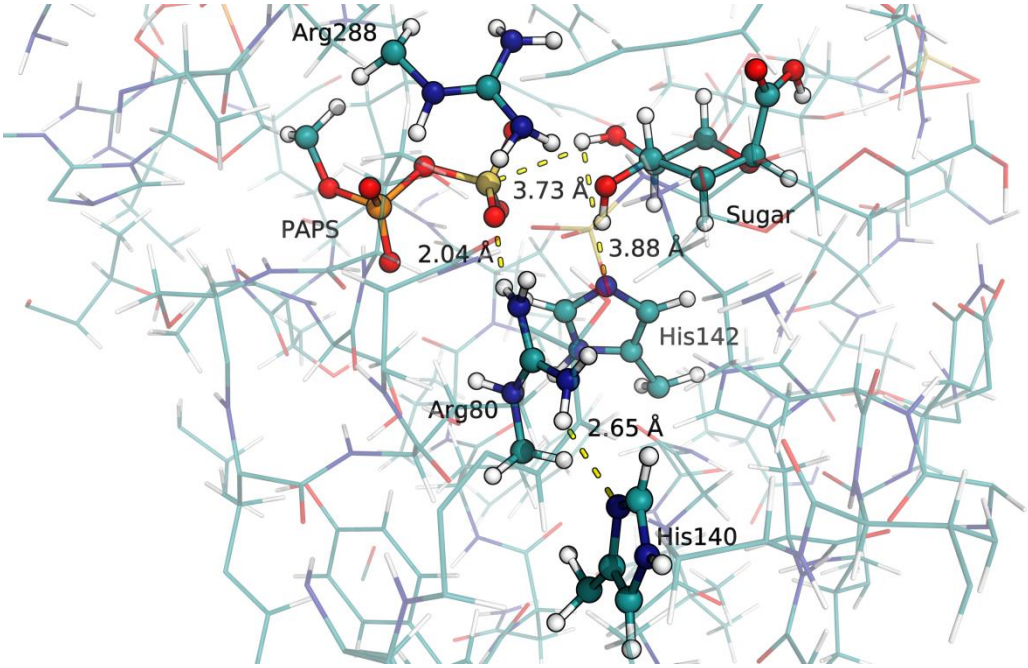

Figure 37 - Resulting structure after optimization at the B3LYP(6-31G(d)):AMBER level and interactions established at the active site.

Having obtained the optimized structure, it is now possible to proceed with the PES scans, similarly to what was previously mentioned. Once again, a PES scan was performed, with the shortening of the S(PAPS)-O2(Sugar) bond as the reaction coordinate. The scan step was 0.05 Å, being shortened to 0.01 Å near the TS. The catalytic mechanism, similarly to 3-OST, likely consists of the transfer of a sulfo group with a concomitant proton transfer, which means it should proceed through a single step. Due to time constraints, only the reactants, transition state and products stationary points have been obtained for this enzyme. The ZPE, thermal and entropic calculations, as well as freely optimized structures, are not being included in this work. The results from this step are evidenced next.

2.2.1. Reactants

In the structure of the reactants, the sulfur atom is positioned at 2.98 Å from the 2-O position on the glucuronic acid unit (O2). The O2 proton has rotated from the position it occupied in the optimized structure, which was mentioned above, and is now hydrogen-bonded to the His142 nitrogen atom (1.97 Å). Arg80 is stabilizing the sulfo group position from PAPS, at 2.13 Å. However, the His140 side chain has drifted apart

from Arg80 and is now at 3.64 Å from it. One interesting thing to note is the somewhat strong hydrogen bond being formed between the phosphate group from PAPS and Arg288 (1.74 Å). This seems to clarify the role of Arg288 as a stabilizer of the PAPS molecule in the active site. Figure 38 shows the main optimized catalytic core and the most important interactions established at the reactants.

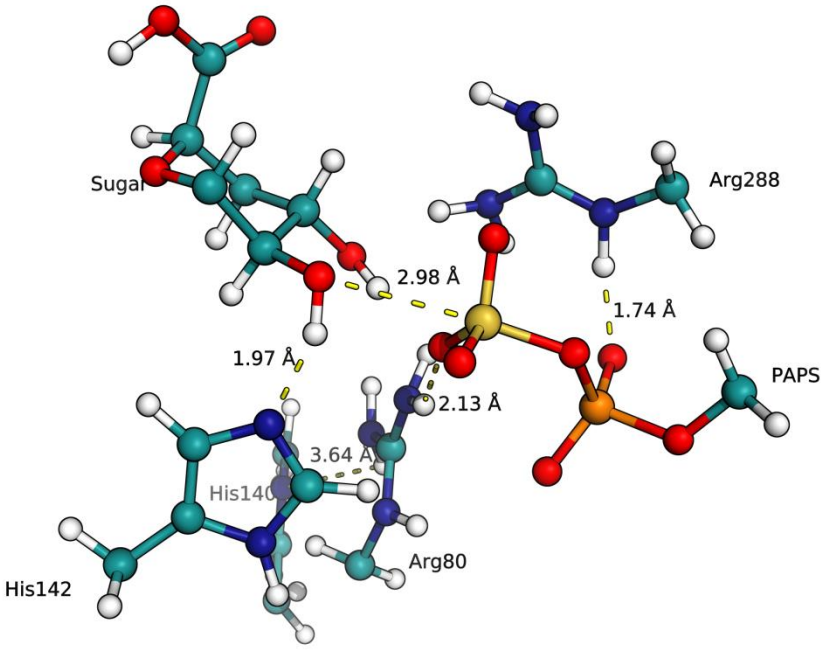

Figure 38 - Representation of the catalytic core and the most important interactions established at the reactants. Distance values are in angstroms.

2.2.2. Transition State

Figure 39 shows the main catalytic core and the most important interactions established at the transition state (TS) structure. The sulfate group is involved in an associative bipyramidal trigonal TS, with the oxygen atom from the sugar and from PAPS in the axial positions, as was previously observed with 3-OST. The group is in an intermediate position between the sugar and PAPS, distanced similarly to both (2.03 and 2.11 Å, respectively). The sugar is now positioned closer to His142, at 1.73 Å. Arg80 is now stabilizing the phosphate group negative charge, which is due to the leaving sulfo group, instead of the sulfo group itself, and is positioned at 1.88 Å from it. His140 remains distanced apart from Arg80, 3.44 Å from it. Surprisingly, Arg288 has formed a very strong hydrogen bond with the phosphate group from PAPS, presumably to, along with Arg80, stabilize its high negative charge.

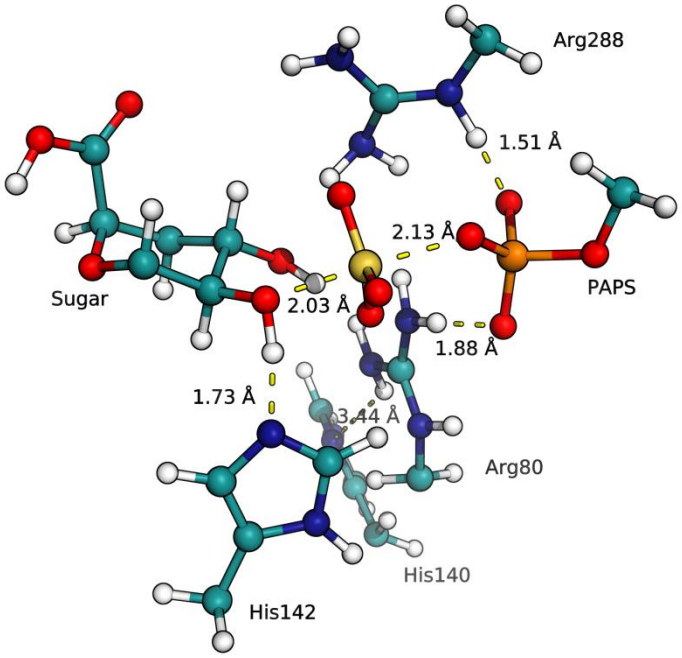

Figure 39 - Transition state (TS) structure, emphasizing the main interactions established. Distance values are in angstroms.

2.2.3. Products

The optimized products geometry is shown in Figure 40. The distance between PAPS and the sugar has increased. The sulfur atom from the transferred sulfate group is positioned 4.75 Å away from the oxygen atom from PAPS to which it was previously bound to, and is now bound to O2 (1.70 Å), thus finishing the catalytic reaction. The proton from O2 has been fully transferred to His142 (1.01 Å) and is now positioned far from the sugar, 3.36 Å away from it. The distance between Arg80 and His140 is now even larger (4.20 Å), whereas the distance between Arg80 and the phosphate group from PAP (which was mentioned in the TS structure description) is now 2.46 Å. One interesting aspect to note, which had not been described nor observed in this or previous works, was the transfer of a proton from the Arg288 amine group to the PAP phosphate group, thus stabilizing it. This suggests that Arg288 has a larger importance to this reaction than was previously thought.

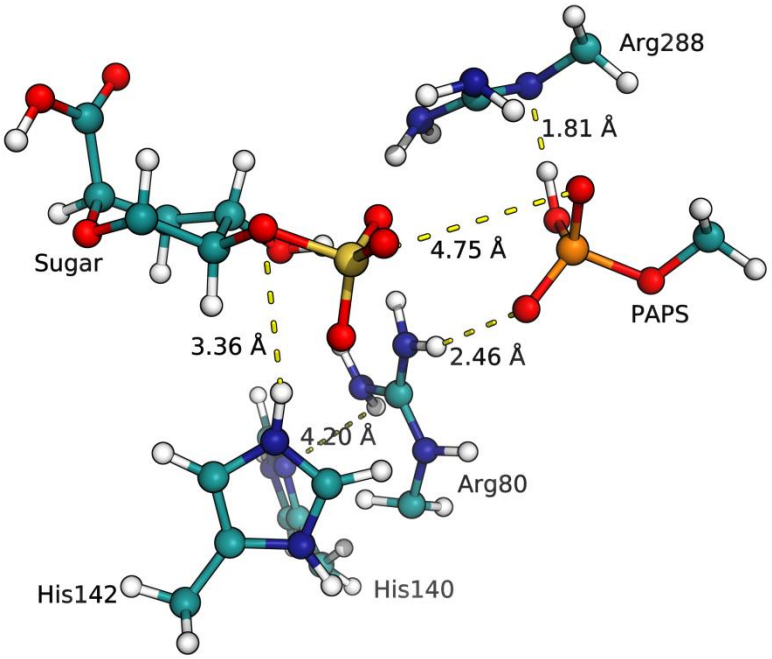

Figure 40 - Representation of the products structure, emphasizing the main interactions established. Distance values are in angstroms.

At the end of this reaction, two residues have a different protonation state from the reactants. Deprotonation of His142 and reprotonation of Arg288 has to take place before the next catalytic cycle begins. This probably occurs after product dissociation and solvation of the active site.

The study of this reaction mechanism has confirmed its similarities with other sulfotransferases, while showing some interesting points. Although there was evidence in the crystallographic structure of the existence of a hydrogen bond between His140 and Arg80, this does not seem to be the case for this enzyme, as was shown in the reactants structure onwards. Furthermore, the Arg288 residue seems to have a very important role in stabilizing the phosphate group from PAPS.

The activation barrier obtained for this reaction was of 35.0 kcal/mol, which is somewhat high. However, there are no experimental data for this enzymatic catalytic mechanism. The products structure shown here seems to have a higher energy than the reactants structure (3.3 kcal/mol difference), which could indicate that this is an endothermic reaction. Both these facts may indicate that the TS structure obtained here does not correspond to either the reactants or the products structure.

Unfortunately, as previously mentioned, the results shown here are not from freely optimized structures, which could alter the results slightly. Furthermore, since the ZPE, thermal and entropic contributions to the free energy were not included, the energetic

results are not entirely accurate. The TS structure is now being optimized, in order to perform IRC calculations in the future. This will allow the clarification of the catalytic mechanism, as well as the evaluation of the validity of the results presently shown.

V. Conclusions

With this work, the catalytic mechanism of both 3-*O*-sulfotransferase and 2-*O*-sulfotransferase, two members of the sulfotransferase family, has been determined through the use of hybrid computational methods, namely the ONIOM method, allied with MD simulations. Through the use of virtual screening, it was also possible to determine several compounds with potential inhibitory activity towards 3-OST. This work is of great importance, not only for the development of new drugs to prevent HSV-1 cell entry, but also in order to further understand the general mechanisms of sulfotransferases, an important enzymatic family.

A model for each enzyme was built, starting from the respective crystallographic structures, with modifications to accommodate for the available computational power, such as the truncation of the saccharides involved in the catalytic reaction. These simplifications have been shown to not significantly affect the obtained results, which are within the expected, structural and energetically for 3-OST; and structurally, since clear energetic results have not been obtained, for 2-OST. The models consisted of 83 QM treated atoms out of 4504 atoms for 3-OST and of 77 QM treated atoms out of 14428 atoms for 2-OST. The residues chosen were based on previous mutagenesis experimental results, which evidenced the crucial residues for the enzyme to possess catalytic activity and the interactions between them and the substrate molecules.

For 3-OST, it was initially thought that the catalytic mechanism would proceed through a charge-relay mechanism, in which Glu184 and His186 would be neutral and Asp189 would be negatively charged. Throughout the reaction, the Glu184 proton would be transferred to His186, coinciding with the transfer of a proton from the sugar to Glu184, and the proton from His186 would be transferred to Asp189. In the end of the catalytic cycle, Asp189 would be the only one with a different protonation state (neutral). However, the results shown here indicate that this is not the case. The mechanism proceeds through a single step, as was expected; however, the protonation state through which it occurs consists of a negatively charged Glu184 and Asp189 and positively charged His186. The reactants structure obtained through optimization of the MD simulations structure shows that Lys215 transfers its proton to Glu184, with this proton being transferred back to Lys215 at the end of the catalytic reaction. Glu184 remains neutral throughout the reaction. The activation free energy obtained at the ONIOM(M06-2X-D3/6-311++G(2d,2p):Amber//B3LYP/6-31G(d):Amber) level is 16.1 kcal/mol, which is consistent with the upper limit determined experimentally for the full cycle (20.4 kcal/mol). A conformational analysis was performed by extracting several structures throughout the MD simulations (eight in total) and by evaluating the

activation free energies obtained for each one. The results clarified the effects of conformational fluctuations in the activation energy of this specific enzyme. Activation free energies ranged from 13.3 kcal/mol to 31.7 kcal/mol, with a weighted average of 14.3 kcal/mol. This value is slightly lower than the activation free energy mentioned above, for the QM/MM calculations at the ONIOM(M06-2X-D3/6-311++G(2d,2p):Amber//B3LYP/6-31G(d):Amber) level. These differences arise from differences in the organization of the active site salt bridges between the phosphate, sulfate and the three lysines. Small differences in this strong interactions network leads to significant shifts in free energies. In order to fully grasp the influence of the conformational space in the activation and reaction free energies, a larger number of structures will need to be extracted from the MD simulations, with the procedure being repeated for each one.

Through the use of virtual screening, using the VsLab VMD plugin, based on the Autodock program in a library containing compounds with known inhibitory activity towards other sulfotransferases and an in-house library of natural compounds, a number of compounds with potential 3-OST inhibitory activity was determined, which is a very important step in the development of new drugs which act on this enzyme. Through analysis of the interactions established at the active site, as well as $\Delta G_{binding}$ values, two compounds, phomoidride B and Barcelonaic acid A, have been identified as potential 3-OST inhibitors. The protocol has been validated and, as such, it can now be applied to a large library of compounds, in order to identify more potential 3-OST inhibitors. Furthermore, several residues have been identified as being important in the binding of both substrate and inhibitors. These include Glu184, His186, Lys215, Lys368, which are also important for the catalytic mechanism, as well as Arg190, Gln255, Arg166 and Lys259, which establish hydrogen bonds with the compounds, stabilizing their position in the active site.

For 2-OST, it was determined that both His140 and His142 were neutral, whereas Arg80 and Arg288 were positively charged. These results fall in line to what was suggested by previous experimental results. This mechanism was believed to proceed through a single step, similarly to 3-OST, where the proton from the 2-OH position in the sugar would be transferred to His142. The other catalytically important residues (Arg80, His140 and Arg288) were believed to only have a spectator role, stabilizing the interactions at the active site level. The results obtained here, however, seem to prove otherwise. Arg288 seems to have a larger importance than what was previously thought, since it transfers a proton to the phosphate group, thus greatly stabilizing its

negative charge left by the leaving of the sulfo group. Moreover, His140 seemed to have a smaller influence in the stabilization of Arg80 than what was previously thought: although it is in a position to stabilize the position of Arg80 at the reactants, during the catalytic reaction it moves away from it. At the end of the catalytic reaction, His142 is doubly protonated and Arg288 is deprotonated. The activation energy obtained for this reaction at the ONIOM(B3LYP/6-31G(d):Amber//B3LYP/6-31G(d):Amber) level was of 35.0 kcal/mol. The reaction energy was 3.3 kcal/mol, which seems to indicate that this reaction is endothermic. However, since the results shown here are not obtained from freely optimized structures and from a small basis set, further work needs to be done to confirm or refute the results shown here.

VI. References

1. P. Chayavichitsilp, J. V. Buckwalter, A. C. Krakowski and S. F. Friedlander, *Pediatrics in Review*, 2009, **30**, 119-130.
2. A. J. Wagstaff, D. Faulds and K. L. Goa, *Drugs*, 1994, **47**, 153-205.
3. G. Darby, H. Field and S. Salisbury, *Nature*, 1981, **289**, 81-83.
4. J. Pottage Jr and H. Kessler, *Infectious agents and disease*, 1995, **4**, 115-124.
5. F. Bevilacqua, A. Marcello, M. Toni, M. Zavattoni, M. Cusini, R. Zerboni, G. Gerna and G. Palù, *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 1991, **4**, 967-969.
6. P. Chrisp and S. P. Clissold, *Drugs*, 1991, **41**, 104-129.
7. M. E. Ganapathy, W. Huang, H. Wang, V. Ganapathy and F. H. Leibach, *Biochemical and biophysical research communications*, 1998, **246**, 470-475.
8. M. R. Harnden, R. L. Jarvest, M. R. Boyd, D. Sutton and R. A. Vere Hodge, *Journal of Medicinal Chemistry*, 1989, **32**, 1738-1743.
9. R. A. V. Hodge, D. Sutton, M. R. Boyd, M. R. Harnden and R. L. Jarvest, *Antimicrobial Agents and Chemotherapy*, 1989, **33**, 1765-1773.
10. D. Earnshaw, T. Bacon, S. Darlison, K. Edmonds, R. Perkins and R. V. Hodge, *Antimicrobial agents and chemotherapy*, 1992, **36**, 2747-2757.
11. D. L. Rabenstein, *Natural product reports*, 2002, **19**, 312-331.
12. J. Turnbull, A. Powell and S. Guimond, *Trends in cell biology*, 2001, **11**, 75-82.
13. R. L. Chen and A. D. Lander, *Journal of Biological Chemistry*, 2001, **276**, 7507-7517.
14. A. Linker, P. Hoffman, P. Sampson and K. Meyer, *Biochimica et Biophysica Acta*, 1958, **29**, 443-444.
15. J. D. Esko and L. Zhang, *Current opinion in structural biology*, 1996, **6**, 663-670.
16. L. Zhang, G. David and J. D. Esko, *Journal of Biological Chemistry*, 1995, **270**, 27127-27135.
17. S. Gulberti, V. Lattard, M. Fondeur, J.-C. Jacquinet, G. Mulliert, P. Netter, J. Magdalou, M. Ouzzine and S. Fournel-Gigleux, *Journal of Biological Chemistry*, 2005, **280**, 1417-1425.
18. J. Kreuger and L. Kjellén, *Journal of Histochemistry & Cytochemistry*, 2012, **60**, 898-907.
19. M. Busse, A. Feta, J. Presto, M. Wilén, M. Grønning, L. Kjellén and M. Kusche-Gullberg, *Journal of Biological Chemistry*, 2007, **282**, 32802-32810.
20. B.-T. Kim, H. Kitagawa, J.-i. Tamura, T. Saito, M. Kusche-Gullberg, U. Lindahl and K. Sugahara, *Proceedings of the National Academy of Sciences*, 2001, **98**, 7176-7181.
21. K. Lidholt and U. Lindahl, *Biochem. J*, 1992, **287**, 21-29.
22. J. M. Whitelock and R. V. Iozzo, *Chemical reviews*, 2005, **105**, 2745-2764.
23. K. Grobe, J. Ledin, M. Ringvall, K. Holmborn, E. Forsberg, J. D. Esko and L. Kjellén, *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2002, **1573**, 209-215.
24. M. A. S. Pinhal, B. Smith, S. Olson, J.-i. Aikawa, K. Kimata and J. D. Esko, *Proceedings of the National Academy of Sciences*, 2001, **98**, 12984-12989.
25. P. Jemth, E. Smeds, A.-T. Do, H. Habuchi, K. Kimata, U. Lindahl and M. Kusche-Gullberg, *Journal of Biological Chemistry*, 2003, **278**, 24371-24376.
26. C. D. O'Donnell, V. Tiwari, M.-J. Oh and D. Shukla, *Virology*, 2006, **346**, 452-459.
27. V. Tiwari, C. D. O'Donnell, M.-J. Oh, T. Valyi-Nagy and D. Shukla, *Biochemical and biophysical research communications*, 2005, **338**, 930-937.
28. D. Xu, V. Tiwari, G. Xia, C. Clement, D. Shukla and J. Liu, *Biochem. J*, 2005, **385**, 451-459.
29. X. Ai, A.-T. Do, M. Kusche-Gullberg, U. Lindahl, K. Lu and C. P. Emerson, *Journal of Biological Chemistry*, 2006, **281**, 4969-4976.

30. F. Gong, P. Jemth, M. L. E. Galvis, I. Vlodavsky, A. Horner, U. Lindahl and J.-p. Li, *Journal of Biological Chemistry*, 2003, **278**, 35152-35158.
31. L.-Å. Fransson, M. Belting, F. Cheng, M. Jönsson, K. Mani and S. Sandgren, *Cellular and Molecular Life Sciences CMLS*, 2004, **61**, 1016-1024.
32. C. Freeman and J. Hopwood, in *Heparin and Related Polysaccharides*, Springer, 1992, pp. 121-134.
33. P. W. Robbins and F. Lipmann, *Journal of the American Chemical Society*, 1956, **78**, 2652-2653.
34. R. Raman, J. Myette, G. Venkataraman, V. Sasisekharan and R. Sasisekharan, *Biochemical and Biophysical Research Communications*, 2002, **290**, 1214-1219.
35. S. C. Edavettal, K. A. Lee, M. Negishi, R. J. Linhardt, J. Liu and L. C. Pedersen, *Journal of Biological Chemistry*, 2004, **279**, 25789-25797.
36. A. F. Moon, S. C. Edavettal, J. M. Krahn, E. M. Munoz, M. Negishi, R. J. Linhardt, J. Liu and L. C. Pedersen, *Journal of Biological Chemistry*, 2004, **279**, 45185-45193.
37. J. Kreuger, M. Salmivirta, L. Sturiale, G. Giménez-Gallego and U. Lindahl, *Journal of Biological Chemistry*, 2001, **276**, 30744-30752.
38. S. L. Bullock, J. M. Fletcher, R. S. Beddington and V. A. Wilson, *Genes & Development*, 1998, **12**, 1894-1906.
39. H. E. Bülow and O. Hobert, *Neuron*, 2004, **41**, 723-736.
40. T. Kinnunen, Z. Huang, J. Townsend, M. M. Gatdula, J. R. Brown, J. D. Esko and J. E. Turnbull, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 1507-1512.
41. H. N. Bethea, D. Xu, J. Liu and L. C. Pedersen, *Proceedings of the National Academy of Sciences*, 2008, **105**, 18724-18729.
42. A. Préchoux, C. Halimi, J.-P. Simorre, H. Lortat-Jacob and C. Laguri, *ACS Chemical Biology*, 2015, **10**, 1064-1071.
43. M. A. Pinhal, B. Smith, S. Olson, J.-i. Aikawa, K. Kimata and J. D. Esko, *Proceedings of the National Academy of Sciences*, 2001, **98**, 12984-12989.
44. H. H. Goldstine and A. Goldstine, *Mathematical Tables and Other Aids to Computation*, 1946, 97-110.
45. E. Schrödinger, *Physical Review*, 1926, **28**, 1049-1070.
46. M. Planck, *Annalen der Physik*, 1901, **309**, 564-566.
47. R. G. Parr and W. Yang, *Density-functional theory of atoms and molecules*, Oxford university press, 1989.
48. L. H. Thomas, *Mathematical Proceedings of the Cambridge Philosophical Society*, 1927, **23**, 542-548.
49. E. Fermi, *Rend. Accad. Naz. Lincei*, 1927, **6**, 32.
50. P. Hohenberg and W. Kohn, *Physical Review*, 1964, **136**, B864-B871.
51. M. Born and P. Jordan, *Z. Physik*, 1925, **34**, 858-888.
52. M. Born and R. Oppenheimer, *Annalen der Physik*, 1927, **389**, 457-484.
53. D. R. Hartree, Mathematical Proceedings of the Cambridge Philosophical Society, 1928.
54. J. C. Slater, *Physical Review*, 1930, **35**, 210.
55. G. G. Hall, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 1951.
56. C. C. J. Roothaan, *Reviews of modern physics*, 1951, **23**, 69.
57. J. A. Pople and G. A. Segal, *The Journal of Chemical Physics*, 1965, **43**, S136-S151.
58. W. Kohn and L. J. Sham, *Physical Review*, 1965, **140**, A1133.
59. S. H. Vosko, L. Wilk and M. Nusair, *Canadian Journal of Physics*, 1980, **58**, 1200-1211.

60. R. Ditchfield, W. J. Hehre and J. A. Pople, *The Journal of Chemical Physics*, 1971, **54**, 724-728.
61. S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Physical Chemistry Chemical Physics*, 2012, **14**, 12431-12441.
62. E. Derat, J. Bouquant and S. Humbel, *Journal of Molecular Structure: THEOCHEM*, 2003, **632**, 61-69.
63. T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery, K. Morokuma and M. J. Frisch, *Journal of Chemical Theory and Computation*, 2006, **2**, 815-826.
64. D. Bakowies and W. Thiel, *The Journal of Physical Chemistry*, 1996, **100**, 10580-10594.
65. F. Maseras and K. Morokuma, *Journal of Computational Chemistry*, 1995, **16**, 1170-1179.
66. S. Humbel, S. Sieber and K. Morokuma, *The Journal of chemical physics*, 1996, **105**, 1959.
67. M. Svensson, S. Humbel, R. D. Froese, T. Matsubara, S. Sieber and K. Morokuma, *The Journal of Physical Chemistry*, 1996, **100**, 19357-19363.
68. S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 15-26.
69. R. Anandakrishnan, B. Aguilar and A. V. Onufriev, H++ 3.2: Web-based computational prediction of protonation states and pK of ionizable groups in macromolecules, http://biophysics.cs.vt.edu/, (accessed 25th September 2015).
70. R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic acids research*, 2012, **40**, W537-W541.
71. J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath and A. Onufriev, *Nucleic Acids Research*, 2005, **33**, W368-W371.
72. J. Myers, G. Grothaus, S. Narayanan and A. Onufriev, *PROTEINS: Structure, Function, and Bioinformatics*, 2006, **63**, 928-938.
73. D. A. Case, T. Darden, T. E. Cheatham III, C. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman and M. Crowley, *University of California, San Francisco*, 2006.
74. W. Humphrey, A. Dalke and K. Schulten, *Journal of molecular graphics*, 1996, **14**, 33-38.
75. N. M. F. S. A. Cerqueira, J. Ribeiro, P. A. Fernandes and M. J. Ramos, *International Journal of Quantum Chemistry*, 2011, **111**, 1208-1212.
76. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235-242.
77. M. e. Frisch, G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci and G. e. Petersson, Gaussian 09, Gaussian, Inc. Wallingford, CT, 2009.
78. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo and T. Lee, *Journal of computational chemistry*, 2003, **24**, 1999-2012.
79. J. Wang, W. Wang, P. A. Kollman and D. A. Case, *Journal of molecular graphics and modelling*, 2006, **25**, 247-260.
80. C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *The Journal of Physical Chemistry*, 1993, **97**, 10269-10280.
81. J. A. Izaguirre, D. P. Catarello, J. M. Wozniak and R. D. Skeel, *The Journal of chemical physics*, 2001, **114**, 2090.
82. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577-8593.
83. K. D. Hammonds and J.-P. Ryckaert, *Computer Physics Communications*, 1991, **62**, 336-351.
84. A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648.

85. C. Lee, W. Yang and R. G. Parr, *Physical Review B*, 1988, **37**, 785-789.
86. A. D. Becke, *The Journal of chemical physics*, 1996, **104**, 1040.
87. Y. Zhao and D. G. Truhlar, *Theoretical Chemistry Accounts*, 2008, **120**, 215-241.
88. A. D. Boese and J. M. Martin, *The Journal of chemical physics*, 2004, **121**, 3405-3416.
89. T. Yanai, D. P. Tew and N. C. Handy, *Chemical Physics Letters*, 2004, **393**, 51-57.
90. A. D. Becke, *The Journal of chemical physics*, 1997, **107**, 8554-8560.
91. H. L. Schmider and A. D. Becke, *The Journal of chemical physics*, 1998, **108**, 9624-9631.
92. J.-D. Chai and M. Head-Gordon, *Physical Chemistry Chemical Physics*, 2008, **10**, 6615-6620.
93. J. P. Perdew, K. Burke and M. Ernzerhof, *Physical review letters*, 1996, **77**, 3865.
94. C. Adamo and V. Barone, *The Journal of chemical physics*, 1999, **110**, 6158-6170.
95. C. Adamo and V. Barone, *The Journal of chemical physics*, 1998, **108**, 664.
96. S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics*, 2010, **132**, -.
97. E. Chapman, S. Ding, P. G. Schultz and C.-H. Wong, *Journal of the American Chemical Society*, 2002, **124**, 14524-14525.
98. M. D. Best, A. Brik, E. Chapman, L. V. Lee, W. C. Cheng and C. H. Wong, *ChemBioChem*, 2004, **5**, 811-819.
99. D. J. Abraham, *Burger's medicinal chemistry and drug discovery*, Wiley Online Library, 2003.
100. D. J. Newman, G. M. Cragg and K. M. Snader, *Journal of natural products*, 2003, **66**, 1022-1037.
101. M. S. Butler, *Natural product reports*, 2005, **22**, 162-195.
102. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *Journal of computational chemistry*, 1998, **19**, 1639-1662.
103. T. U. Consortium, *Nucleic Acids Research*, 2015, **43**, D204-D212.
104. E. Chapman, M. D. Best, S. R. Hanson and C. H. Wong, *Angewandte Chemie - International Edition*, 2004, **43**, 3526-3548.
105. W. Min, B. P. English, G. Luo, B. J. Cherayil, S. Kou and X. S. Xie, *Accounts of Chemical Research*, 2005, **38**, 923-931.
106. R. D. Smiley and G. G. Hammes, *Chemical reviews*, 2006, **106**, 3080-3094.
107. C. Chen, M. E. Layton, S. M. Sheehan and M. D. Shair, *Journal of the American Chemical Society*, 2000, **122**, 7424-7425.
108. T. T. Dabrah, T. Kaneko, W. Massefski and E. B. Whipple, *Journal of the American Chemical Society*, 1997, **119**, 1594-1598.
109. H. Jayasuriya, R. G. Ball, D. L. Zink, J. L. Smith, M. A. Goetz, R. G. Jenkins, M. Nallin-Omstead, K. C. Silverman, G. F. Bills and R. B. Lingham, *Journal of natural products*, 1995, **58**, 986-991.
110. H. Van De Waterbeemd and E. Gifford, *Nature reviews Drug discovery*, 2003, **2**, 192-204.
111. M. P. Gleeson, *Journal of medicinal chemistry*, 2008, **51**, 817-834.
112. J. Hodgson, *Nature Biotechnology*, 2001, **19**, 722-726.

Appendix A

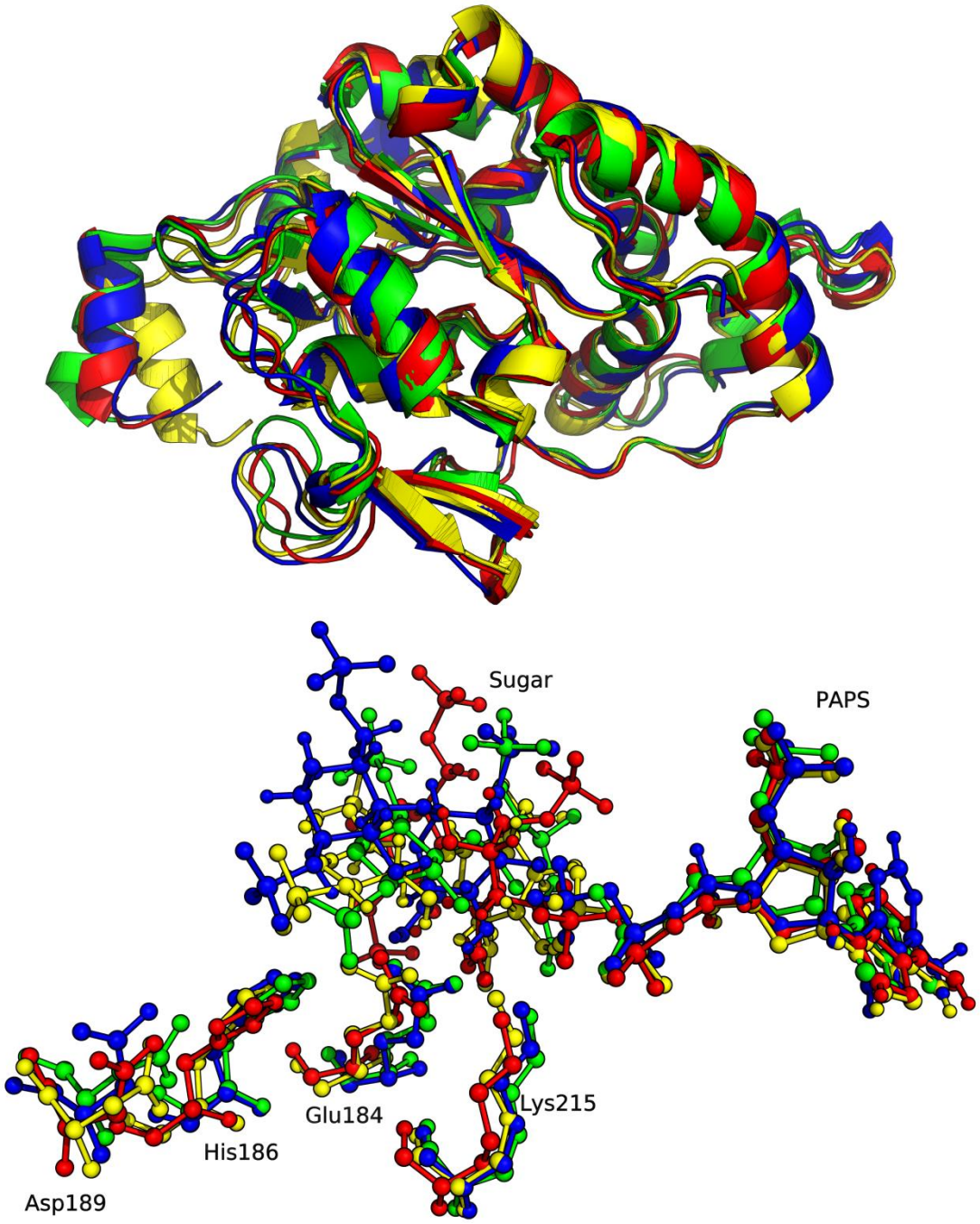

Figure 41 - Representation of the superimposition of the resulting structures extracted from the MD simulations for 3-OST (Red - model A; Blue - Model B; Yellow - Model C) and the crystallographic structure (Green), with a cartoon representation for the whole protein (top) and a CPK representation of the key active site residues and substrate molecules (bottom).